

DUALITY APPROACH TO BILEVEL PROGRAMS WITH A CONVEX LOWER LEVEL*

ANIL ASWANI[†] AND AURÉLIEN OUATTARA[†]

Abstract. Bilevel programs can be reformulated as a single-level program by replacing the lower level problem with an optimality condition. This paper studies the optimality condition of upper-bounding the lower level objective by a new dual function constructed to be differentiable (unlike the usual Lagrangian dual function), which leads to a duality-based reformulation (with elements of regularization) of the bilevel program. We study the existence of solutions, relationship between local solutions of the bilevel program and its reformulation, constraint qualification, stability and continuity of solutions, and convergence of stationary points as the regularization is decreased. In particular, we show that under appropriate regularity (characterized by a sufficient condition) of the solution mapping of the lower level problem there is convergence to true stationary points of the unregularized problem. These results show approximate bilevel programs and their duality-based reformulations do not share the same pathologies that occur for bilevel programs or their reformulations. This is used to argue that, from a modeling perspective, approximate bilevel programs may be a more appropriate model for many applications. We conclude with two numerical examples to demonstrate how our duality-based approach can be used to solve practical bilevel programs.

Key words. duality, regularization, variational analysis, bilevel programming

AMS subject classifications. 90C30, 90C31, 90C46

1. Introduction. Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ be vectors, and consider the following (optimistic) bilevel programming problem:

$$\begin{aligned} \min_{x,y} \quad & F(x, y) \\ \text{BLP} \quad & \text{s.t. } G(x) \leq 0 \\ & y \in \arg \min_y \{f(x, y) \mid g(x, y) \leq 0\} \end{aligned}$$

where F, f are scalar-valued and G, g are vector-valued functions. (We preclude equality constraints for expositional convenience, and our results in this paper generalize to the setting with equality constraints.) If we call x the upper-level decision variables and y the lower-level decision variables, then we can consider $\min_y \{f(x, y) \mid g(x, y) \leq 0\}$ to be the lower level problem. In this paper, we will specifically study the situation where the lower level problem is convex (i.e., f, g are convex in y for fixed x).

The solution approach for BLP is to replace the lower level problem by some optimality conditions and then solve the reformulated problem. One method [2, 17, 27] replaces the lower level problem with the KKT conditions, which yields a mathematical program with equilibrium constraints (MPEC) that can be solved using corresponding algorithms. Another method [30, 45] replaces the lower level problem by the inequalities $f(x, y) \leq \varphi(x)$ and $g(x, y) \leq 0$, where $\varphi(x) = \min_y \{f(x, y) \mid g(x, y) \leq 0\}$ is the value function. The intuition is that any y satisfying these inequalities gives an objective function value that is less than or equal to the minimum.

However, the KKT and value function methods face a number of difficulties. Both the KKT and value function methods generate a formulation that does not satisfy constraint qualification [39, 40, 45]. For regularized KKT methods, stationary points of

***Funding:** This work was supported in part by NSF Award CMMI-1450963 and the Philippine-California Advanced Research Institutes (PCARI).

[†]Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA (aaswani@berkeley.edu, aurelien.ouattara@berkeley.edu).

the regularized problems converge to *weaker* stationary points of the original bilevel program as the regularization decreases [19, 24, 40, 42]. The value function method introduces a non-differentiable constraint $f(x, y) - \varphi(x) \leq 0$ (since the value function is not differentiable), and so numerical optimization requires specialized descent algorithms that provide smoothing of the value function [25].

This paper develops a duality-based approach to solving bilevel programs where the lower level is convex. Our resulting reformulation of BLP is such that (i) each term is differentiable, (ii) constraint qualification holds after regularization, and (iii) stationary points of the regularized reformulation converge to *true* stationary points of BLP. These features allow numerical solution of our reformulation (and BLP) using standard algorithms and software for nonlinear optimization. We also provide new results on the continuity and stability of bilevel programs and approximate bilevel programs. These results are used to argue that from a modeling perspective, approximate bilevel programs may be a more useful or appropriate model for many settings.

1.1. Background. The idea of the duality-based approach is to pose the optimality condition of the lower level problem as the inequalities $f(x, y) \leq h(\lambda, x)$, $\lambda \geq 0$, and $g(x, y) \leq 0$, where $h(\lambda, x)$ is a dual function. Under conditions implying zero duality gap, these constraints force y to be a minimizer of the lower level problem. We previously proposed a duality-based approach in a paper on inverse optimization with noisy data [3], though the prior formulation is not differentiable because of the use of Lagrangian dual functions. This paper constructs an alternative dual function that is differentiable under mild conditions. We also study constraint qualification and convergence of stationary points, which were not previously considered in [3].

The KKT reformulation [2, 17, 27] (and its regularized versions [24, 40, 42]) is differentiable when the defining functions are twice differentiable, while the value function approach [30, 45] leads to a reformulation that is not differentiable. This prevents use of standard optimization software, though specialized algorithms have been designed [25, 30]. Our current work generalizes the value function approach by lifting the value function with additional variables (i.e., the Lagrange multipliers $\lambda \geq 0$) to ensure differentiability and enable use of standard optimization solvers.

These reformulations can sometimes be relaxed to ensure constraint qualification. For the KKT approach, regularization occurs by smoothing complementary constraints, but constraint qualification holds under difficult-to-check preconditions [24, 40, 42]. For the value function approach, the reformulation can be relaxed with $f(x, y) \leq \varphi(x) + \epsilon$ to allow for ϵ -solutions, but constraint qualification holds under difficult-to-check preconditions [25]. Here, we propose consideration of ϵ -solutions in the sense of [28, 29], meaning we relax the constraints to $g(x, y) \leq \epsilon$. This subtle difference provides enough structure to show constraint qualification holds for regularized versions of our duality-based approach without difficult-to-check preconditions.

A related question is if stationary points of the regularized problems converge to stationary points of the original bilevel program. For KKT methods, convergence occurs towards weaker stationary points [19, 24, 40, 42]. For the value function approach, limiting stationary points are well-behaved for *calm* bilevel programs [25, 45]; but calmness can be a restrictive assumption. For our duality-based approach, stationary points of regularized problems will converge to true stationary points of the unregularized problem when the solution mapping of the lower level problem satisfies a regularity condition that we characterize with a sufficient condition. We also show that local minima of the unregularized reformulation match local minima of the original bilevel program under conditions satisfied by many practical bilevel programs.

Lastly, it is known that solutions to bilevel programs are neither continuous nor stable because the lower level problem is not continuous with respect to the upper-level decision variables [15]. Thus it has been proposed to consider ϵ -solutions (i.e., $f(x, y) \leq \varphi(x) + \epsilon$) of the lower level problem [15, 23, 25, 26], which leads to asymptotically continuous solutions [15, 26]. Continuity with respect to a parametrization of the bilevel program holds under difficult-to-check preconditions [23]. We show that ϵ -solutions in the sense of [28, 29] (i.e., $g(x, y) \leq \epsilon$) lead to continuity of the bilevel program with respect to parametrizations and without difficult-to-check preconditions.

1.2. Outline. Section 2 provides preliminaries, including our assumptions about BLP and four new results in variational analysis about regularity that may be of independent interest. Section 3 defines a new dual function whose maximizers are equivalent to those of the Lagrangian dual function. Our dual is differentiable, unlike the Lagrangian dual (which is only directionally differentiable). We use our dual to define a duality-based reformulation (DBP) of BLP in section 4, and the equivalence of DBP and BLP is proved. Next, we consider constraint qualification and consistency of approximation of regularized versions of DBP. Section 5 studies the stability and continuity of solutions of approximate BLP, which is used to argue that approximate BLP may be a more appropriate model in certain settings. Last, section 6 presents two numerical examples to show how DBP can be used to solve practical bilevel programs.

2. Preliminaries. This section describes our notation, gives useful definitions from variational analysis [36], and states our assumptions about BLP. We also include four new results in variational analysis that may be of independent interest, including: a generalization of the boundedness theorem to set-valued mappings, a uniform regularity for a specific class of lower- \mathcal{C}^2 functions, a uniform regularity for constraint sets formed by this class, and a partial characterization of the normal cone of the convex superset of a convex set. These new results substantially simplify later proofs.

2.1. Notation. Let $\|\cdot\|$ denote the ℓ_2 norm, and $\langle \cdot, \cdot \rangle$ is the dot product. The ball with radius r centered at x is $\mathcal{B}(x, r) = \{x' : \|x' - x\| \leq r\}$. We use $(\subset, \supset) \subseteq, \supseteq$ to indicate (proper) subsets and supersets, respectively. The domain of a function f is $\text{dom}(f)$, and we will assume functions are extended real-valued. The set \mathcal{C}^2 consists of all twice continuously differentiable functions. The *prime* notation is used in two different ways, the meaning of which is made clear from the context: We use A' to denote the transpose of a matrix A , and we use x' to distinguish a variable from x .

The upper and inner limit of a function f at \bar{x} are written as $\limsup_{x \rightarrow \bar{x}} f(x)$ and $\liminf_{x \rightarrow \bar{x}} f(x)$, respectively. Recall f is lower semicontinuous (lsc) at \bar{x} if $\liminf_{x \rightarrow \bar{x}} f(x) \geq f(\bar{x})$. The outer and inner limit of a set-valued mapping S at \bar{x} are denoted $\limsup_{x \rightarrow \bar{x}} S(x)$ and $\liminf_{x \rightarrow \bar{x}} S(x)$, respectively. Recall S is outer semicontinuous (osc) at \bar{x} if $\limsup_{x \rightarrow \bar{x}} S(x) \subseteq S(\bar{x})$, inner semicontinuous (isc) at \bar{x} if $\liminf_{x \rightarrow \bar{x}} S(x) \supseteq S(\bar{x})$, and continuous at \bar{x} if it is both osc and isc at \bar{x} .

Our first new result generalizes the boundedness theorem to set-valued mappings. Because of the technical peculiarities of continuity for set-valued mappings, we require additional assumptions (i.e., pointwise convexity and boundedness) beyond continuity.

LEMMA 1. *Let X be a compact set, and consider a set-valued mapping $S(x)$ that is convex-valued, continuous, and bounded for each $x \in X$. Then $S(X)$ is bounded.*

Proof. Suppose the set $S(X)$ is not bounded. Then there exist sequences $x^\nu \in X$ and $s^\nu \in S(x^\nu)$ such that $\|s^\nu\| \rightarrow \infty$. Since X is compact, there is some convergent subsequence by the Bolzano-Weierstrass theorem; and so by extracting this subsequence we can assume $x^\nu \rightarrow \bar{x}$ for some $\bar{x} \in X$. Now consider the sequence $s^\nu / \|s^\nu\|$;

note the norm of each term is 1. Hence there is some convergent subsequence, and so by extracting this subsequence we can assume $s^\nu/\|s^\nu\| \rightarrow w$ for some $w \neq 0$. Next choose any $t \in S(\bar{x})$, and note that by continuity of S there exists $t^\nu \in S(x^\nu)$ such that $t^\nu \rightarrow t$. For any $\tau \geq 0$, there is a ν large enough such that $\tau/\|s^\nu\| < 1$. And because S is convex-valued, this means we have $(1 - \tau/\|s^\nu\|) \cdot t^\nu + \tau/\|s^\nu\| \cdot s^\nu \in S(x^\nu)$ for ν large enough. Taking the limit, we have $t + \tau w \in S(\bar{x})$. But this is a contradiction because: $w \neq 0$, $\tau \geq 0$ is arbitrary, and S is bounded (by assumption) at $\bar{x} \in X$. Thus, we have shown by contradiction that $S(X)$ is bounded. \square

The outer and inner limit of sets C_t are denoted $\limsup_{t \rightarrow \bar{t}} C_t$ and $\liminf_{t \rightarrow \bar{t}} C_t$, respectively. If the outer and inner limits agree, then the limit: $\lim_t C_t = \limsup_t C_t = \liminf_t C_t$ exists. Let f_t, f be single-valued functions. We say f_t epi-converges to f (or equivalently $\text{e-lim}_{t \rightarrow \bar{t}} f_t = f$) if $\lim_{t \rightarrow \bar{t}} \{(x, u) : u \geq f_t(x)\} = \{(x, u) : u \geq f(x)\}$. Let S_t, S be multi-valued functions. The graphical outer and inner limits are denoted $\text{g-lim sup}_t S_t$ and $\text{g-lim inf}_t S_t$, respectively. If $\text{g-lim sup}_t S_t = \text{g-lim inf}_t S_t$, then the graphical limit $\text{g-lim } S_t$ exists and is given by: $\text{g-lim } S_t = \text{g-lim sup } S_t = \text{g-lim inf } S_t$.

2.2. Normal Cones. Let C be a set. Then $\text{co}(C)$ is the convex hull of C , and the indicator function $\delta_C(x)$ of C is defined as: $\delta_C(x) = 0$ if $x \in C$, and $\delta_C(x) = +\infty$ if $x \notin C$. $\hat{N}_C(x)$ is the regular normal cone of C at x , and recall $v \in \hat{N}_C(x)$ if $\langle v, x' - x \rangle \leq o(\|x' - x\|)$ for all $x' \in C$. The normal cone of C at x is $N_C(x)$, and note $v \in N_C(x)$ if there are sequences $x^\nu \rightarrow x$ with $x^\nu \in C$, and $v^\nu \rightarrow v$ with $\hat{N}_C(x^\nu)$. In this paper, we will usually have $N_C(x) = \hat{N}_C(x)$; and so the regular normal cone notation $\hat{N}_C(x)$ will only be used when needed to distinguish from $N_C(x)$.

The non-negative orthant $\Lambda = \{\lambda : \lambda \geq 0\}$ is a closed, convex set; and its (regular) normal cone is $N_\Lambda(\lambda) = \{x \leq 0 : \lambda_i x_i = 0\}$. The non-positive orthant $\Upsilon = \{x : x \leq 0\}$ is a closed, convex set; and its (regular) normal cone is $N_\Upsilon(x) = \{\lambda \geq 0 : \lambda_i x_i = 0\}$.

The following new result relates the normal cone of a convex constraint set to the normal cone of a convex superset of the constraint set.

PROPOSITION 2. *Suppose **A1** and **R1** hold, and let X, Z be compact, convex sets with $Z \supseteq \{y \in \phi(x) : x \in X\}$ and $\phi(x) = \{y : g(x, y) \leq 0\}$. If $y \in \phi(x)$, then we have $N_Z(y) \subseteq \{\nabla_y g(x, y)' \lambda : \lambda \in N_\Lambda(g(x, y))\}$.*

Proof. $N_Z(y) = \{v : \langle v, w - y \rangle \leq 0, \forall w \in Z\}$ since Z is convex, by Theorem 6.9 in [36]. But $\phi(x) \subseteq Z$ by definition, and so for any $v \in N_Z(y)$ we have $\langle v, w - y \rangle \leq 0$ for all $w \in \phi(x)$. This means $N_{\phi(x)}(y) \supseteq N_Z(y)$. But $N_{\phi(x)}(y)$ is Clarke regular at y by Theorem 6.9 of [36] since it is closed due to **A1**. Thus, $N_{\phi(x)}(y) = \{\nabla_y g(x, y)' \lambda : \lambda \in N_\Lambda(g(x, y))\}$ because **R1** allows us to apply Theorem 6.14 of [36]. \square

2.3. Constraint Qualification. It is useful to define generalized derivatives of a function f that is finite at \bar{x} . A vector v is a regular subgradient, written $v \in \hat{\partial} f(\bar{x})$, if $f(x) \geq f(\bar{x}) + \langle v, x - \bar{x} \rangle + o(\|x - \bar{x}\|)$. A vector v is a subgradient, written $v \in \partial f(\bar{x})$, if there are sequences $x^\nu \rightarrow \bar{x}$ and $v^\nu \in \hat{\partial} f(x^\nu)$ with $f(x^\nu) \rightarrow f(\bar{x})$ and $v^\nu \rightarrow v$. A vector v is a horizon subgradient, written $v \in \partial^\infty f(\bar{x})$, if there are sequences $x^\nu \rightarrow \bar{x}$, $\lambda^\nu \rightarrow 0$, and $v^\nu \in \hat{\partial} f(x^\nu)$ with $f(x^\nu) \rightarrow f(\bar{x})$ and $\lambda^\nu v^\nu \rightarrow v$. For the functions we consider, the set of subgradients will usually be equal to the set of regular subgradients.

These definitions can be used to define constraint qualification for nondifferentiable constraint sets [18]. For a constraint set $\mathcal{C} = \{g(x) \leq 0\}$, consider any point $\bar{x} \in \mathcal{C}$ and define the set of active constraints to be $I = \{i : g_i(\bar{x}) = 0\}$. This constraint set satisfies linear independence constraint qualification (LICQ) if all choices of $v_i \in \partial g_i(\bar{x})$, for all $i \in I$, are linearly independent. This constraint set satisfies the

Mangasarian-Fromovitz constraint qualification (MFCQ) if there is a d such that for all choices of $v_i \in \partial g_i(\bar{x})$, for all $i \in I$, we have $\langle v_i, d \rangle < 0$. When $g(x)$ is differentiable, these definitions simplify to the usual ones for constraint qualification.

2.4. Regularity. We will need to consider several notions of regularity for functions, and we define these regularity notions below

1. A function f is lower- \mathcal{C}^2 if for each $\bar{x} \in \text{dom}(f)$ there is some neighborhood V such that the function can be written as $f(x) = \max_y \{g(x, y) \mid y \in Y\}$; where Y is a compact set, and $g(x, y) \in \mathcal{C}^2$ on $V \times Y$. [34, 36]
2. An lsc function f is primal-lower-nice (pln) at $\bar{x} \in \text{dom}(f)$ if there exist $r, c, \tau > 0$ so that if $t \geq \tau$, $\|v\| \leq ct$, $\|x - \bar{x}\| \leq r$, and $v \in \partial f(x)$; then $f(x') \geq f(x) + \langle v, x' - x \rangle - (t/2)\|x' - x\|^2$, for all x' with $\|x' - x\| \leq r$. We say f is pln if it is pln at all $\bar{x} \in \text{dom}(f)$. [22, 33]
3. An lsc function f is prox-regular (pr) at $\bar{x} \in \text{dom}(f)$ for $\bar{v} \in \partial f(\bar{x})$ if there exist $r > 0$ and $t \geq 0$ such that $f(x') \geq f(x) + \langle v, x' - x \rangle - (t/2)\|x' - x\|^2$, for all x' with $\|x' - \bar{x}\| \leq r$, when $v \in \partial f(x)$, $\|v - \bar{v}\| < r$, $\|x - \bar{x}\| < r$, and $f(x) < f(\bar{x}) + r$. If f is pr at \bar{x} for all $\bar{v} \in \partial f(\bar{x})$, then it is pr at \bar{x} . We say f is pr if it is pr at all $\bar{x} \in \text{dom}(f)$. [32, 36]

A lower- \mathcal{C}^2 function is pln [22, 33], and a pln function is pr [32, 36]. One class of lower- \mathcal{C}^2 functions has additional regularity that is useful for proving later results.

LEMMA 3. *Consider a lower- \mathcal{C}^2 function given by $f(x) = \max_y \{g(x, y) \mid y \in Y\}$ for all x . If X is a compact set, then $t = \max_{x,y} \{\|D_x^2 g(x, y)\| \mid x \in X, y \in Y\} \geq 0$ is finite and such that for all $x', x \in X$ and $v \in \partial f(x)$ we have $f(x') \geq f(x) + \langle v, x' - x \rangle - (t/2)\|x' - x\|^2$. Moreover, $c = \max_{x,y} \{\|\nabla_x g(x, y)\| \mid x \in X, y \in Y\} \geq 0$ is finite and such that $\|v\| \leq c$ for all $v \in \partial f(x)$ with $x \in X$.*

Proof. Let $\sigma(x) = \arg \max_y \{g(x, y) \mid y \in Y\}$, and consider any $\bar{y} \in \sigma(x)$. Taylor's theorem implies $f(x') - f(x) \geq g(x', \bar{y}) - g(x, \bar{y}) \geq \langle \nabla_x g(x, \bar{y}), x' - x \rangle - (t/2)\|x' - x\|^2$ with $t = \max_{x,y} \{\|D_x^2 g(x, y)\| \mid x \in X, y \in Y\}$. This t is finite by Example 1.11 of [36] because $g(x, y) \in \mathcal{C}^2$ by assumption, and X, Y are compact. Since the above inequality holds for any $\bar{y} \in \sigma(x)$, the first part of the result follows by noting Theorem 2.1 of [14] shows the subgradient of f is $\partial f(x) = \text{co}\{\nabla_x g(x, \bar{y}) : \bar{y} \in \sigma(x)\}$. The second part of the result follows by recalling the equation for the subgradient of f , and noting $c = \max_{x,y} \{\|\nabla_x g(x, y)\| \mid x \in X, y \in Y\}$ is finite by the same argument as earlier. \square

The regularity notions of pln and pr defined above can be extended to a set C by applying the corresponding definitions to the indicator function δ_C for the set. Our next result proves prox-regularity of constraint sets defined by the class of lower- \mathcal{C}^2 functions from Lemma 3, and it partly generalizes existing results from [11, 32] on the prox-regularity of constraint sets defined by functions with Lipschitz derivatives.

LEMMA 4. *Consider the lower- \mathcal{C}^2 function given by $f(x) = \max_y \{g(x, y) \mid y \in Y\}$ for all x . If $f(x)$ satisfies MFCQ, then $C = \{x : f(x) \leq 0\}$ is prox-regular.*

Proof. Theorem 2.1 of [14] shows the subgradient of f_i is $\partial f_i(x) = \text{co}\{\nabla_x g_i(x, \bar{y}_i) : \bar{y}_i \in \sigma_i(x)\}$, where $\sigma_i(x) = \arg \max_y \{g_i(x, y) \mid y \in Y\}$. Since a lower- \mathcal{C}^2 function is locally Lipschitz continuous [34, 36], this means $\sum \lambda_i f_i(x)$ is locally Lipschitz continuous for $\lambda \geq 0$. So the horizon subgradient of $\sum \lambda_i f_i(x)$ is $\{0\}$ by Theorem 9.13.b of [36], which with MFCQ means we can apply Theorem 10.49 of [36]: This means the subgradient of the indicator function for C , given by $\delta_C(x) = [\delta_Y \circ f](x)$, is $\partial \delta_C(x) = \{\sum \lambda_i \nabla_x g_i(x, \bar{y}_i) : \bar{y}_i \in \sigma_i(x), \lambda \in N_Y(f(x))\}$. Define $\Xi(x, v) = \{\lambda \in N_Y(f(x)) : v = \sum \lambda_i \nabla_x g_i(x, \bar{y}_i), \bar{y}_i \in \sigma_i(x)\}$, and let $\bar{v} \in \partial \delta_C(\bar{x})$ for any $\bar{x} \in C$. We show there exists

$e > 0$ such that $\Pi(\bar{x}, \bar{v}) = \{\lambda \in \Xi(x, v) : \|x - \bar{x}\| < e, \|v - \bar{v}\| < e, x \in C\}$ is bounded. Supposing it is not bounded, there exist sequences $x^\nu \rightarrow x$ with $x^\nu \in C$, $v^\nu \rightarrow v$ with $v^\nu \in \partial\delta_C(x^\nu)$, $\bar{y}_i^\nu \in \sigma_i(x^\nu)$, and $\lambda^\nu \in \Xi(x^\nu, v^\nu)$ such that $\|\lambda^\nu\| \rightarrow \infty$. Note we have

$$(1) \quad v^\nu = \sum \lambda_i^\nu \nabla_x g_i(x^\nu, \bar{y}_i^\nu) \text{ and } \lambda_i^\nu f_i(x^\nu) = 0.$$

Since $\lambda^\nu / \|\lambda^\nu\| = 1$ and $\bar{y}_i^\nu \in \sigma_i(x^\nu) \subseteq Y$, there is some convergent subsequence by the Bolzano-Weierstrass theorem; and so by extracting this subsequence we can assume: $\bar{y}_i^\nu \rightarrow \bar{y}_i$ for some $\bar{y}_i \in \sigma_i(x)$ (since σ_i is osc by the Berge maximum theorem [7]), and $\lambda^\nu / \|\lambda^\nu\| \rightarrow \lambda$ for some $\lambda \neq 0$ with $\lambda \geq 0$ (since N_Γ is a cone). Dividing (1) by $\|\lambda^\nu\|$ and taking the limit gives $0 = \sum \lambda_i \nabla_x g_i(x, \bar{y}_i)$ and $\lambda_i f_i(x) = 0$, where we have used the twice continuous differentiability of g and the Lipschitz continuity of f . But this contradicts the MFCQ, and so we conclude there exists $e > 0$ such that $\Pi(\bar{x}, \bar{v})$ is bounded. Let $\pi_{\bar{x}, \bar{v}}$ be a bound on the norm of elements in $\Pi(\bar{x}, \bar{v})$. Next note δ_Γ is convex, and so for $x, x' \in C$ we have $\langle \lambda, f(x') - f(x) \rangle \leq 0$ for any $\lambda \in N_\Gamma(f(x))$. And since f satisfies the hypothesis of Lemma 3, this means for any $\bar{x} \in C$ and $\bar{v} \in \partial\delta_C(\bar{x})$ there exists $e > 0$ such that $\langle \sum \lambda_i \nabla_x g_i(x, \bar{y}_i), x' - x \rangle \leq (\pi_{x, v} \cdot t) / 2 \|x' - x\|^2$, for all $x' \in C$ such that $\|x' - \bar{x}\| \leq e$, when $v \in \partial\delta_C(x)$, $\|v - \bar{v}\| < e$, and $\|x - \bar{x}\| < e$. Here we have $\bar{y}_i \in \sigma_i(x)$ and $t = \max_i t_i$, where the t_i are as given in Lemma 3 for f_i and $X_i = \{x : \|x - \bar{x}\| \leq e\}$. But $\sum \lambda_i \nabla_x g_i(x, \bar{y}_i) \in \partial\delta_C(x)$, and we have that C is closed since $f(x)$ is Lipschitz continuous. Thus the result follows because we have shown C satisfies the definition of prox-regularity. \square

We will also need equivariant versions of these regularity notions, which can be defined for both functions and sets. In particular, we say that a family of functions (sets) is equi-primal-lower-nice at \bar{x} if each function (set) in the family is primal-lower-nice at \bar{x} with the same constants r, c, τ (from the definition). Similarly, a family of functions (sets) is equi-prox-regular at \bar{x} if each function (set) in the family is prox-regular at \bar{x} with the same constants r, t (from the definition).

Using these notions, we can prove an equivariant version of Lemma 4 for parametric lower- \mathcal{C}^2 functions from the class defined in Lemma 3. The key assumption needed is structural independence of the perturbation from the variable of the function.

LEMMA 5. *Let U be a compact set, and consider a parametric family of (multivariate) lower- \mathcal{C}^2 functions given by $f(x, u) = \max_y \{g(x, y) + h(y, u) \mid y \in Y\}$ for all x and $u \in U$. If $g, h \in \mathcal{C}^2$ and $f(x, u)$ satisfies MFCQ for each $u \in U$, then $C(u) = \{x : f(x, u) \leq 0\}$ is equi-prox-regular for $u \in U$.*

Proof. Observe that $C(u)$ is osc for $u \in U$ by Example 5.8 of [36] because $f(x, u)$ is locally Lipschitz continuous since it is lower- \mathcal{C}^2 [34, 36]. Next note the indicator function for $C(u)$ is given by $\delta_{C(u)}(x) = [\delta_\Gamma \circ f(\cdot, u)](x)$, and that Theorem 2.1 of [14] shows the subgradient of $f_i(\cdot, u)$ is $\partial f_i(x, u) = \text{co}\{\nabla_x g_i(x, \bar{y}_i) : \bar{y}_i \in \sigma_i(x, u)\}$, where $\sigma_i(x, u) = \arg \max_y \{g_i(x, y) + h_i(y, u) \mid y \in Y\}$. Since a lower- \mathcal{C}^2 function is locally Lipschitz continuous [34, 36], this means the sum $\sum \lambda_i f_i(x, u)$ is also locally Lipschitz continuous for $\lambda \geq 0$. So the horizon subgradient of $\sum \lambda_i f_i(x, u)$ is $\{0\}$ by Theorem 9.13.b of [36]. This along with MFCQ means we can apply Theorem 10.49 of [36], which implies the subgradient of $\delta_{C(u)}$ is $\partial\delta_{C(u)}(x) = \left\{ \sum \lambda_i \nabla_x g_i(x, \bar{y}_i) : \bar{y}_i \in \sigma_i(x, u), \lambda \in N_\Gamma(f(x, u)) \right\}$. Now define $\Xi(x, v) = \{\lambda \in N_\Gamma(f(x, u)) : v = \sum \lambda_i \nabla_x g_i(x, \bar{y}_i), \bar{y}_i \in \sigma_i(x, u)\}$, and let $\bar{v} \in \partial\delta_{C(u)}(\bar{x})$ for any $\bar{x} \in C(u)$ and $u \in U$. We show there exists $e > 0$ such that $\Pi(\bar{x}, \bar{v}) = \{\lambda \in \Xi(x, v) : \|x - \bar{x}\| < e, \|v - \bar{v}\| < e, x \in C(u), u \in U\}$ is bounded. Supposing it is not bounded, then there exist sequences $u^\nu \rightarrow u$ with $u^\nu \in U$, $x^\nu \rightarrow x$ with $x^\nu \in C(u^\nu)$ and $x \in C(u)$, $v^\nu \rightarrow v$ with $v^\nu \in \partial\delta_C(x^\nu, u^\nu)$,

$\bar{y}_i^\nu \in \sigma_i(x^\nu, u^\nu)$, and $\lambda^\nu \in \Xi(x^\nu, v^\nu)$ such that $\|\lambda^\nu\| \rightarrow \infty$. Thus we have

$$(2) \quad v^\nu = \sum \lambda_i^\nu \nabla_x g_i(x^\nu, \bar{y}_i^\nu) \text{ and } \lambda_i^\nu f_i(x^\nu, u^\nu) = 0.$$

Since $\lambda^\nu / \|\lambda^\nu\| = 1$ and $\bar{y}_i^\nu \in \sigma_i(x^\nu) \subseteq Y$, there is some convergent subsequence by the Bolzano-Weierstrass theorem; and so by extracting this subsequence we can assume: $\bar{y}_i^\nu \rightarrow \bar{y}_i$ for some $\bar{y}_i \in \sigma_i(x, u)$ (since σ_i is osc by the maximum theorem of Berge [7]), and $\lambda^\nu / \|\lambda^\nu\| \rightarrow \lambda$ for some $\lambda \neq 0$ with $\lambda \geq 0$ (since N_Υ is a cone). Dividing (2) by $\|\lambda^\nu\|$ and taking the limit gives $0 = \sum \lambda_i \nabla_x g_i(x, \bar{y}_i)$ and $\lambda_i f_i(x) = 0$, where we have used the twice continuous differentiability of g and the Lipschitz continuity of f . But this contradicts the MFCQ, and so we conclude there exists $e > 0$ such that $\Pi(\bar{x}, \bar{v})$ is bounded. Let $\pi_{\bar{x}, \bar{v}}$ be a bound on the norm of elements in $\Pi(\bar{x}, \bar{v})$. Next note δ_Υ is convex, and so for $x, x' \in C(u)$ we have $\langle \lambda, f(x', u) - f(x, u) \rangle \leq 0$ for any $\lambda \in N_\Upsilon(f(x, u))$. And since f satisfies the hypothesis of Lemma 3, this means for any $\bar{x} \in C(u)$ and $\bar{v} \in \partial \delta_{C(u)}(\bar{x})$ there exists $e > 0$ such that $\langle \sum \lambda_i \nabla_x g_i(x, \bar{y}_i), x' - x \rangle \leq (\pi_{x, v} \cdot t)/2 \|x' - x\|^2$ for all $x' \in C(u)$ such that $\|x' - \bar{x}\| \leq e$, when $v \in \partial \delta_{C(u)}(x)$, $\|v - \bar{v}\| < e$, and $\|x - \bar{x}\| < e$. Here we have $\bar{y}_i \in \sigma_i(x, u)$ and $t = \max_i t_i$, where the t_i are as given in Lemma 3 for f_i and $X_i = \{x : \|x - \bar{x}\| \leq e\}$. But $\sum \lambda_i \nabla_x g_i(x, \bar{y}_i) \in \partial \delta_{C(u)}(x, u)$, and we have that $C(u)$ is closed for each $u \in U$ since $f(\cdot, u)$ is Lipschitz continuous. This shows $C(u)$ for each $u \in U$ satisfies the definition of prox-regularity. However, careful inspection shows $e, \pi_{x, v}, t_i, t$ are independent of u . And so we conclude that $C(u)$ is equi-prox-regular for $u \in U$. \square

2.5. Assumptions. For the lower level problem, we define its value function $\varphi(x) = \min_y \{f(x, y) \mid g(x, y) \leq 0\}$, solution set $s(x) = \arg \min_y \{f(x, y) \mid g(x, y) \leq 0\}$, and feasible set $\phi(x) = \{y : g(x, y) \leq 0\}$. The Lagrangian dual function (LDF) is $\psi(\lambda, x) = \inf_y f(x, y) + \langle \lambda, g(x, y) \rangle$. For the BLP, let $X = \{x : G(x) \leq 0\}$.

We also define some assumptions about and regularity conditions for BLP. It is important to note that not all assumptions and regularity conditions are used in every result, and that our purpose in listing all the conditions here is to aggregate them into a single location. Our first set of assumptions relate to the lower level problem.

A1. The $f(x, y), g(x, y)$ are convex in y (for fixed x) and satisfy $f, g \in \mathcal{C}^2$.

A2. There exists a compact, convex set $Y \supset \{y : \exists x \in X \text{ s.t. } g(x, y) \leq 0\}$.

R1. For each $x \in X$, there exists y such that $g(x, y) < 0$.

These conditions ensure the lower level problem and its Lagrange dual problem are solvable, meaning the minimum (maximum) is attained and the set of optimal solutions is nonempty and compact. Also, the pointwise nature of **R1** will be important for proving solvability of BLP.

PROPOSITION 6. *Suppose **A1**, **A2** and **R1** hold. Then $\arg \max_\lambda \{\psi(\lambda, x) \mid \lambda \geq 0\}$ and $\arg \min_y \{f(x, y) \mid g(x, y) \leq 0\}$ are compact and nonempty.*

Proof. The result follows from Example 1.11 of [36] and Theorem 2.165 of [10]. \square

Our next assumptions concern BLP, and they ensure smoothness in the objective function of BLP and regularity in the constraints $G(x) \leq 0$. These conditions, when combined with the previous conditions, ensure BLP is solvable (to be shown later).

A3. The $F(x, y), G(x)$ are twice continuously differentiable in x, y .

R2. The set X is compact and nonempty, and $G(x)$ satisfies MFCQ for each $x \in X$.

3. Constrained Lagrangian Dual Function. The numerical issue with the Lagrangian dual function (LDF) $\psi(\lambda, x) = \inf_y f(x, y) + \langle \lambda, g(x, y) \rangle$, is that it is generally nondifferentiable in λ .

Example 7. The example of linear programming is classical: Let $A \in \mathbb{R}^{p \times m}$, $b \in \mathbb{R}^p$, $c \in \mathbb{R}^m$, and define $f(x, y) = \langle c, y \rangle$ and $g(x, y) = Ax - b$. Then, the LDF is

$$(3) \quad \psi(\lambda, x) = \begin{cases} -\langle b, \lambda \rangle, & \text{if } A'\lambda = -c \text{ and } \lambda \geq 0 \\ -\infty, & \text{otherwise} \end{cases}$$

For λ_0 such that $A'\lambda_0 = -c$ and $\lambda_0 \geq 0$, this LDF is directionally differentiable in directions d such that $A'd = 0$ and $\lambda_0 + td \geq 0$ for $t > 0$ small enough. However, this LDF is not differentiable because it is discontinuous in directions d such that $A'd \neq 0$ or $\lambda_0 + td \not\geq 0$ for any $t > 0$. \diamond

Because the LDF is generally not differentiable, this limits its utility in reformulating bilevel programs because in general closed-form expressions for the domain of the LDF are not available. In this section, we construct an alternative dual function that is designed to be differentiable while retaining the saddle point and strong duality properties of the LDF.

3.1. Definition. We define the Constrained Lagrangian Dual Function (CDF):

$$(4) \quad h(\lambda, x) = \min_y \{f(x, y) + \langle \lambda, g(x, y) \rangle \mid y \in Y\}.$$

The difference as compared to the (classical) LDF is the domain of minimization of the Lagrangian $\mathcal{L}(x, y, \lambda) = f(x, y) + \langle \lambda, g(x, y) \rangle$. The LDF is defined as the infimum of the Lagrangian over the entire space \mathbb{R}^m . In contrast, the CDF is defined as the minimum of the Lagrangian over a compact, convex set Y that is a proper superset of $\{y : \exists x \in X \text{ s.t. } g(x, y) \leq 0\}$.

An important feature of the CDF is it maintains the strong duality of the LDF, and its solutions are a saddle point to the Lagrangian $\mathcal{L}(x, y, \lambda)$. Our first result establishes an equivalence between solutions of the CDF and LDF.

THEOREM 8. *Suppose **A1**, **A2** and **R1** hold. Then $\arg \max_{\lambda} \{\psi(\lambda, x) \mid \lambda \geq 0\}$ is non-empty and compact, $\max_{\lambda} \{h(\lambda, x) \mid \lambda \geq 0\} = \max_{\lambda} \{\psi(\lambda, x) \mid \lambda \geq 0\}$, and $\arg \max_{\lambda} \{h(\lambda, x) \mid \lambda \geq 0\} = \arg \max_{\lambda} \{\psi(\lambda, x) \mid \lambda \geq 0\}$*

Proof. For the optimization problem $\min_y \{f(x, y) \mid g(x, y) \leq 0\}$, we can associate (as in Example 11.46 of [36]) the generalized Lagrangian $l(x, y, \lambda) = f(x, y) + \langle \lambda, g(x, y) \rangle - \delta_{\Lambda}(\lambda)$. Note $\psi(\lambda, x) = \inf_y l(x, y, \lambda)$ for $\lambda \geq 0$, and that because f, g are differentiable by **A1** we have $\partial_y l(x, y, \lambda) = \nabla_y f(x, y) + \nabla_y g(x, y)' \lambda$ and $\partial_{\lambda}[-l](x, y, \lambda) = -g(x, y) + N_{\Lambda}(\lambda)$. Since $s(x)$ is compact and nonempty by **Proposition 6**, let $y^* \in s(x)$. Theorem 11.50 and Corollary 11.51 of [36] give that (i) $\lambda^* \in \arg \max_{\lambda} \{\psi(\lambda, x) \mid \lambda \geq 0\}$ exists, (ii) $\psi(\lambda^*, x) = f(x, y^*)$, and (iii) $0 \in \partial_y l(x, y^*, \lambda^*)$ and $0 \in \partial_{\lambda}[-l](x, y^*, \lambda^*)$. Our next step is to associate a generalized Lagrangian to the optimization problem $\min_{y \in Y} \{f(x, y) \mid g(x, y) \leq 0\}$. Following Example 11.46 of [36], the corresponding generalized Lagrangian is given by $\ell(x, y, \lambda) = \delta_Y(y) + f(x, y) + \langle \lambda, g(x, y) \rangle - \delta_{\Lambda}(\lambda)$. Note $h(\lambda, x) = \min_y \{\ell(x, y, \lambda) \mid y \in Y\}$ for $\lambda \geq 0$, and that $\partial_y \ell(x, y, \lambda) = N_Y(y) + \nabla_y f(x, y) + \nabla_y g(x, y)' \lambda$ and $\partial_{\lambda}[-\ell](x, y, \lambda) = -g(x, y) + N_{\Lambda}(\lambda)$. Since $0 \in N_Y(y)$ and $0 \in \partial_y l(x, y^*, \lambda^*)$, this implies $0 \in \partial_y \ell(x, y^*, \lambda^*)$. Similarly, $0 \in \partial_{\lambda}[-l](x, y^*, \lambda^*)$ yields $0 \in \partial_{\lambda}[-\ell](x, y^*, \lambda^*)$. Thus, we can apply Theorem 11.50 and Corollary 11.51 of [36], which gives (i) $\lambda^* \in \arg \max_{\lambda} \{h(\lambda, x) \mid \lambda \geq 0\}$, and (ii) $h(\lambda^*, x) = f(x, y^*)$. And so $h(\lambda^*, x) = \psi(\lambda^*, x)$, proving the first part of the result.

For the second result, recall from above that $\lambda^* \in \arg \max_{\lambda} \{h(\lambda, x) \mid \lambda \geq 0\}$. This means $\arg \max_{\lambda} \{h(\lambda, x) \mid \lambda \geq 0\} \supseteq \arg \max_{\lambda} \{\psi(\lambda, x) \mid \lambda \geq 0\}$. Next, observe Theorem 11.50 and Corollary 11.51 of [36] give that (i) $\mu^* \in \arg \max_{\lambda} \{h(\lambda, x) \mid \lambda \geq 0\}$ exists, (ii) $h(\mu^*, x) = f(x, y^*)$, and (iii) $0 \in \partial_y \ell(x, y^*, \mu^*)$ and $0 \in \partial_{\lambda}[-\ell](x, y^*, \mu^*)$. The condition $0 \in \partial_{\lambda}[-\ell](x, y^*, \mu^*)$ implies both $0 \in \partial_{\lambda}[-l](x, y^*, \mu^*)$ and $y \in \phi(x)$. This second consequence implies $y \in \text{int}(Y)$ by **A2**, and hence $N_Y(y) = \{0\}$. As a result, we have $0 \in \partial_y l(x, y^*, \mu^*)$ because $0 \in \partial_y \ell(x, y^*, \mu^*)$. Applying Theorem 11.50 and Corollary 11.51 of [36] leads to the conclusion that $\mu^* \in \arg \max_{\lambda} \{\psi(\lambda, x) \mid \lambda \geq 0\}$. Thus, we have $\arg \max_{\lambda} \{h(\lambda, x) \mid \lambda \geq 0\} \subseteq \arg \max_{\lambda} \{\psi(\lambda, x) \mid \lambda \geq 0\}$. Since we have shown both set inclusions, this implies equality and hence the second result. \square

This result is nontrivial because a slight (and subtle) relaxation of the hypothesis causes the result to become untrue. In particular, suppose we replace **A2** with an assumption on the existence of a compact, convex set Z that satisfies $Z \supseteq \{y : \exists x \in X \text{ s.t. } g(x, y) \leq 0\}$. (The difference from **A2** is that Y is a proper superset, while Z is a superset, of $\{y : \exists x \in X \text{ s.t. } g(x, y) \leq 0\}$.) The above result fails because in general we have $\arg \max_{\lambda} \{\eta(\lambda, x) \mid \lambda \geq 0\} \supseteq \arg \max_{\lambda} \{\psi(\lambda, x) \mid \lambda \geq 0\}$ for $\eta(\lambda, x) = \min_y \{f(x, y) + \langle \lambda, g(x, y) \rangle \mid y \in Z\}$. The following example provides one situation where this superset is proper, and this emphasizes the importance of **A2**.

Example 9. Consider the problem: $f(x, y) = y$, $g_1(x, y) = -y - 1$, and $g_2(x, y) = y - 1$. If $Z = \phi(x) = \{y : y \in [-1, 1]\}$, then $\eta(\lambda, x) = -|1 - \lambda_1 + \lambda_2| - \lambda_1 - \lambda_2$ and

$$(5) \quad \psi(\lambda, x) = \begin{cases} -\lambda_1 - \lambda_2, & \text{if } -\lambda_1 + \lambda_2 = -1 \text{ and } \lambda \geq 0 \\ -\infty, & \text{otherwise} \end{cases}$$

Simple calculations give $\arg \max_{\lambda} \{\psi(\lambda, x) \mid \lambda \geq 0\} = \{\lambda : \lambda_1 = 1 \text{ and } \lambda_2 = 0\}$, which is singleton, and $\arg \max_{\lambda} \{\eta(\lambda, x) \mid \lambda \geq 0\} = \{\lambda : \lambda_1 \in [0, 1] \text{ and } \lambda_2 = 0\}$, which is *not* singleton. And so we have $\arg \max_{\lambda} \{\eta(\lambda, x) \mid \lambda \geq 0\} \supset \arg \max_{\lambda} \{\psi(\lambda, x) \mid \lambda \geq 0\}$. \diamond

More formally, this complication when $Z \supseteq \{y : \exists x \in X \text{ s.t. } g(x, y) \leq 0\}$ is explained by **Proposition 2**. If $y^* \in s(x)$ and $\lambda^* \in \arg \max \{\eta(\lambda, x) \mid \lambda \geq 0\}$, then

$$(6) \quad \begin{aligned} 0 &\in N_Z(y^*) + \nabla_y f(x, y^*) + \nabla_y g(x, y^*)' \lambda^* \\ 0 &\in -g(x, y^*) + N_{\Lambda}(\lambda^*) \end{aligned}$$

When **Proposition 2** holds, we have $N_Z(y) \subseteq \{\nabla_y g(x, y)' \lambda : 0 \in -g(x, y) + N_{\Lambda}(\lambda)\}$ (where we used the equivalence of $\lambda \in N_{\Lambda}(g(x, y))$ and $0 \in -g(x, y) + N_{\Lambda}(\lambda)$). The issue is that $N_Z(y)$ has essentially the same form as $\nabla_y g(x, y)' \lambda$, and so the number of λ^* satisfying (6) can increase. However, when $Y \supset \{y : \exists x \in X \text{ s.t. } g(x, y) \leq 0\}$ we have $N_Y(y^*) = \{0\}$ and so the number of λ^* is preserved.

Because the CDF is constructed to have the same solutions as the LDF, the CDF enjoys the same strong duality and saddle point properties of the LDF.

COROLLARY 10. *Suppose **A1**, **A2** and **R1** hold. If $\lambda^* \in \arg \max_{\lambda} \{h(\lambda, x) \mid \lambda \geq 0\}$ and $y^* \in (x)$, then 1. $\min_y \{f(x, y) \mid g(x, y) \leq 0\} = \max_{\lambda} \{h(\lambda, x) \mid \lambda \geq 0\} = \mathcal{L}(x, y^*, \lambda^*)$ and 2. $\mathcal{L}(x, y^*, \lambda) \leq \mathcal{L}(x, y^*, \lambda^*) \leq \mathcal{L}(x, y, \lambda^*)$ for all $y \in \mathbb{R}^m$ and $\lambda \geq 0$.*

Proof. **Theorem 8** implies $\arg \max_{\lambda} \{h(\lambda, x) \mid \lambda \geq 0\} = \arg \max_{\lambda} \{\psi(\lambda, x) \mid \lambda \geq 0\}$, so $\lambda^* \in \arg \max_{\lambda} \{\psi(\lambda, x) \mid \lambda \geq 0\}$. Theorem 11.50 and Corollary 11.51 of [36] give $\min_y \{f(x, y) \mid g(x, y) \leq 0\} = \max_{\lambda} \{h(\lambda, x) \mid \lambda \geq 0\} = l(x, y^*, \lambda^*)$ and $l(x, y^*, \lambda) \leq l(x, y^*, \lambda^*) \leq l(x, y, \lambda^*)$ for all $y \in \mathbb{R}^m$ and $\lambda \geq 0$, where $l(x, y, \lambda)$ is the generalized Lagrangian in the proof of **Theorem 8**. But $\mathcal{L}(x, y, \lambda) = l(x, y, \lambda)$ when $\lambda \geq 0$. \square

A2 is again a crucial assumption, and the result does not hold if this assumption is relaxed to the set Z defined above. Recall the difference from **A2** is that Y is a proper superset, while Z is a superset, of $\{y : \exists x \in X \text{ s.t. } g(x, y) \leq 0\}$. The saddle point result (i.e., the second part of the corollary) fails for \mathcal{L} . (However, a saddle point result holds for the generalized Lagrangian $\ell(x, y, \lambda)$ defined in the proof of [Theorem 8](#).) The following continuation of the previous example shows this.

Example 9 (continued). Recall the problem with $f(x, y) = y$, $g_1(x, y) = -y - 1$, and $g_2(x, y) = y - 1$. If $Y = \{y : y \in [-2, 2]\}$ and $Z = \phi(x) = \{y : y \in [-1, 1]\}$, then $\eta(\lambda, x) = -|1 - \lambda_1 + \lambda_2| - \lambda_1 - \lambda_2$ and $\arg \max_{\lambda} \{\eta(\lambda, x) \mid \lambda \geq 0\} = \{\lambda : \lambda_1 \in [0, 1] \text{ and } \lambda_2 = 0\}$. Choosing $y = -2$ and $\lambda^* = 0$ gives $\mathcal{L}(x, y, \lambda^*) = -2 < \mathcal{L}(x, y^*, \lambda^*) = -1$ because $y^* = -1$. So maximizers of $\eta(\lambda, x)$ (which uses Z) do not satisfy the saddle point property for \mathcal{L} . In contrast, note $h(\lambda, x) = -2 \cdot |1 - \lambda_1 + \lambda_2| - \lambda_1 - \lambda_2$. A simple calculation gives $\arg \max_{\lambda} \{h(\lambda, x) \mid \lambda \geq 0\} = \{\lambda : \lambda_1 = 1 \text{ and } \lambda_2 = 0\}$ (matching [Theorem 8](#) since $\arg \max_{\lambda} \{\psi(\lambda, x) \mid \lambda \geq 0\} = \{\lambda : \lambda_1 = 1 \text{ and } \lambda_2 = 0\}$). Thus, for the solution provided by $h(\lambda, x)$ we get $\mathcal{L}(x, y, \lambda^*) = -1 \geq \mathcal{L}(x, y^*, \lambda^*) = -1 \geq \mathcal{L}(x, y^*, \lambda) = -1 - 2\lambda_2$ for all $y \in \mathbb{R}^m$ and $\lambda \geq 0$, which matches [Corollary 10](#). \diamond

3.2. Differentiability. The distinguishing property of the CDF is that it is differentiable, while the LDF is only directionally differentiable (see [Example 7](#)). The differentiability occurs because the CDF is defined as a minimization over a compact set that is independent of λ, x . Such independence between the domain of minimization and these variables is important, since otherwise we would not have differentiability. This explains why the value function $\varphi(x)$ remains non-differentiable even when the feasible set is compact – because the constraints may be a function of x .

The differentiability of the CDF is essentially a restatement of [Theorem 4.13](#) and [Remark 4.14](#) of [\[10\]](#). In particular, if we define $\sigma(\lambda, x) = \arg \min_y \{f(x, y) + \langle \lambda, g(x, y) \rangle \mid y \in Y\}$, then we can state the differentiability of the CDF.

THEOREM 11. *Suppose **A1**, **A2** and **R1** hold. If (λ, x) is such that $\sigma(\lambda, x)$ is singleton; then the CDF is differentiable at (λ, x) , and its gradient is given by*

$$(7) \quad \begin{aligned} \nabla_x h(\lambda, x) &= \nabla_x f(x, \bar{y}) + \langle \lambda, \nabla_x g(x, \bar{y}) \rangle \\ \nabla_\lambda h(\lambda, x) &= g(x, \bar{y}) \end{aligned}$$

where we have $\{\bar{y}\} = \sigma(\lambda, x) = \arg \min_y \{f(x, y) + \langle \lambda, g(x, y) \rangle \mid y \in Y\}$.

We specifically constructed $h(\lambda, x)$ so as to satisfy the conditions of [Theorem 4.13](#) and [Remark 4.14](#) of [\[10\]](#). Though determining if $\sigma(\lambda, x)$ is singleton can be difficult, there is a simple-to-check condition that ensures this is always the case:

COROLLARY 12. *Suppose **A1**, **A2** and **R1** hold. If $\lambda \geq 0$ and $f(x, y)$ is strictly convex in y for every $x \in X$; then the CDF is differentiable at (λ, x) , and its gradient is given by (7), where we have $\{\bar{y}\} = \sigma(\lambda, x) = \arg \min_y \{f(x, y) + \langle \lambda, g(x, y) \rangle \mid y \in Y\}$.*

Proof. Since $\lambda \geq 0$, we have that $f(x, y) + \langle \lambda, g(x, y) \rangle$ is strictly convex in y for every $x \in X$ (see for instance [Exercise 2.18](#) in [\[36\]](#)). Combining [Example 1.11](#) and [Theorem 2.6](#) of [\[36\]](#) gives that $\sigma(\lambda, x)$ is singleton. We can then apply [Theorem 11](#). \square

For the case where $f(x, y)$ is not strictly convex, we can define a regularized CDF that is guaranteed to be differentiable under a mild condition. In particular, we define the regularized constrained Lagrangian dual function (RDF) to be

$$(8) \quad h_\mu(\lambda, x) = \min_y \{\mu \|y\|^2 + f(x, y) + \langle \lambda, g(x, y) \rangle \mid y \in Y\},$$

where $\mu \geq 0$. We can interpret this as the CDF for an optimization problem where the objective has been changed to $\mu\|y\|^2 + f(x, y)$. The benefit of adding the $\mu\|y\|^2$ term is it makes the objective of the optimization problem defining $h_\mu(\lambda, x)$ strictly convex, and therefore ensures the RDF is differentiable as long as $\mu > 0$. More formally, if $\sigma_\mu(\lambda, x) = \arg \min_y \{\mu\|y\|^2 + f(x, y) + \langle \lambda, g(x, y) \rangle \mid y \in Y\}$, then we have the following:

COROLLARY 13. *Suppose **A1**, **A2** and **R1** hold. If $\lambda \geq 0$ and $\mu > 0$; then the RDF is differentiable at (λ, x) , and its gradient is given by*

$$(9) \quad \begin{aligned} \nabla_x h_\mu(\lambda, x) &= \nabla_x f(x, \bar{y}) + \langle \lambda, \nabla_x g(x, \bar{y}) \rangle \\ \nabla_\lambda h_\mu(\lambda, x) &= g(x, \bar{y}) \end{aligned}$$

where we have $\{\bar{y}\} = \sigma_\mu(\lambda, x) = \arg \min_y \{\mu\|y\|^2 + f(x, y) + \langle \lambda, g(x, y) \rangle \mid y \in Y\}$.

Proof. Since $\|y\|^2$ is strictly convex and $f(x, y)$ is convex, $\mu\|y\|^2 + f(x, y)$ is strictly convex in y for every $x \in X$ (Exercise 2.18 in [36]). So we can apply [Corollary 12](#). \square

More generally, both the CDF and RDF have a strong type of regularity because of their construction. This regularity will be useful for proving subsequent results.

PROPOSITION 14. *Suppose **A1**, **A2** and **R1** hold. Then for any $\mu \geq 0$, we have that $[-h]_\mu(\lambda, x)$ is lower- \mathcal{C}^2 (and hence locally Lipschitz continuous, primal-lower-nice, and prox-regular). Its subgradient is nonempty, compact, and given by*

$$(10) \quad \begin{aligned} \partial_x [-h]_\mu(\lambda, x) &= -\text{co}\{\nabla_x f(x, \bar{y}) + \langle \lambda, \nabla_x g(x, \bar{y}) \rangle \mid \bar{y} \in \sigma_\mu(\lambda, x)\} \\ \partial_\lambda [-h]_\mu(\lambda, x) &= -\text{co}\{g(x, \bar{y}) \mid \bar{y} \in \sigma_\mu(\lambda, x)\} \end{aligned}$$

where $\sigma_\mu(\lambda, x) = \arg \min_y \{\mu\|y\|^2 + f(x, y) + \langle \lambda, g(x, y) \rangle \mid y \in Y\}$.

Proof. We have $[-h]_\mu(\lambda, x) = \max_y \{-\mu\|y\|^2 - f(x, y) - \langle \lambda, g(x, y) \rangle \mid y \in Y\}$, by rewriting the definition of $h_\mu(\lambda, x)$. And so $[-h]_\mu$ is lower- \mathcal{C}^2 by definition. This implies local Lipschitz continuity [34, 36], the primal-lower-nice property [22, 33], and prox-regularity [32, 36]. Theorem 9.13 of [36] gives nonemptiness and compactness of the subgradient, and the formula for the subgradient is due to Theorem 2.1 of [14]. \square

3.3. Convergence Properties. An important aspect of the RDF is it epi-converges to the CDF as $\mu \rightarrow 0$. We begin with a result on this epi-convergence, along with a monotonicity property of the RDF. Note this result does not require $\sigma(\lambda, x)$ to be singleton, and hence applies even when $f(x, y)$ is not strictly convex in y for every $x \in X$. Also, note the epi-convergence result applies to $-h(\lambda, \mu)$ and $-h_\mu(\lambda, \mu)$ since we are typically concerned with maximizing the dual function.

PROPOSITION 15. *Suppose **A1**, **A2** and **R1** hold. Then $[-h]_\mu(\lambda, x)$ is strictly decreasing in μ , and we have that $\text{e-lim}_{\mu \rightarrow 0} [-h]_\mu(\lambda, x) = [-h](\lambda, x)$.*

Proof. The Berge maximum theorem [7] implies $h(\lambda, x)$ and $h_\mu(\lambda, x)$ are continuous (for each fixed $\mu > 0$). Second, note Proposition 7.4.c of [36] gives that for fixed λ, x we have $\text{e-lim}_{\mu \rightarrow 0} \mu\|y\|^2 + f(x, y) + \langle \lambda, g(x, y) \rangle = f(x, y) + \langle \lambda, g(x, y) \rangle$. And so by Theorem 7.33 of [36], we have for fixed λ, x that $\lim_{\mu \rightarrow 0} h_\mu(\lambda, x) = h(\lambda, x)$. Now let $\bar{y} \in \sigma_\mu(\lambda, x)$, and observe that for $0 \leq \mu_1 < \mu_2$ we have $h_{\mu_1}(\lambda, x) \leq \mu_1\|\bar{y}\|^2 + f(x, \bar{y}) + \langle \lambda, g(x, \bar{y}) \rangle < \mu_2\|\bar{y}\|^2 + f(x, \bar{y}) + \langle \lambda, g(x, \bar{y}) \rangle = h_{\mu_2}(\lambda, x)$. Thus, $[-h]_\mu(\lambda, x)$ is strictly decreasing in μ . This implies $\sup_{\mu > 0} [-h]_\mu(\lambda, x) = [-h](\lambda, x)$ since from above we have $\lim_{\mu \rightarrow 0} [-h]_\mu(\lambda, x) = [-h](\lambda, x)$. Finally, using Proposition 7.4.d of [36] gives the desired result: $\text{e-lim}_{\mu \rightarrow 0} [-h]_\mu(\lambda, x) = [-h](\lambda, x)$. \square

Gradient consistency [5, 12, 13, 33, 47] is a stronger notion than convergence of functions, and it relates to the convergence of (generalized) gradients for a convergent sequence of functions. For example [47], the function $\zeta_n(x) = \frac{1}{n} \sin(nx)$ epi-converges to $\zeta(x) = 0$, while $\nabla_x \zeta_n(x) = \cos(nx)$ does not converge to $\nabla_x \zeta(x) = 0$. An additional regularity is needed to ensure gradient consistency holds. In particular, we require a type of equivariability for the family of functions $\{h_\mu(\lambda, x) : \mu > 0\}$. The following result shows that the RDF forms a family of functions that is equi-primal-lower-nice.

PROPOSITION 16. *Suppose **A1**, **A2** and **R1** hold. Then $[-h]_\mu(\lambda, x)$ are equi-primal-lower-nice.*

Proof. Fix $\mu \geq 0$, choose any $(\underline{x}, \underline{\lambda})$, and set $\Sigma = \mathcal{B}((\underline{x}, \underline{\lambda}), 1)$. Note $[-h]_\mu(\lambda, x)$ is lower- \mathcal{C}^2 and satisfies the assumptions in **Lemma 3**. So there is a finite constant $t = \max_{x, y, \lambda} \{\|D_{(x, \lambda)}^2 \mathcal{L}(x, y, \lambda)\| \mid (x, \lambda) \in \Sigma, y \in Y\} \geq 0$ such that for any $(x', \lambda'), (x, \lambda) \in \Sigma$ and $v \in \partial[-h]_\mu(\lambda, x)$ we have $[-h]_\mu(\lambda', x') \geq [-h]_\mu(\lambda, x) + \langle v_x, x' - x \rangle + \langle v_y, \lambda' - \lambda \rangle - (t/2)\|(\lambda', x') - (\lambda, x)\|^2$. Moreover, $c = \max_{x, y, \lambda} \{\|\nabla_{(x, \lambda)} \mathcal{L}(x, y, \lambda)\| \mid (x, \lambda) \in \Sigma, y \in Y\} \geq 0$ is finite and such that $\|v\| \leq c$ for all $v \in \partial[-h]_\mu(\lambda, x)$ with $(x, \lambda) \in \Sigma$. Since $(\underline{x}, \underline{\lambda})$ is arbitrary, $[-h]_\mu(\lambda, x)$ is primal-lower-nice with constants (from the definition) of $r = 1$, c , and $\tau = \max\{1, t\}$. But these constants are independent of μ , and so we get that $[-h]_\mu(\lambda, x)$ is equi-primal-lower-nice. \square

A convergent sequence of equi-primal-lower-nice family of functions is sufficient to show gradient consistency under some additional boundedness. We use the above result to show the subgradients of RDF converge to the subgradient of CDF as $\mu \rightarrow 0$.

THEOREM 17. *Suppose **A1**, **A2** and **R1** hold. Then we have gradient consistency $\partial_{(x, \lambda)}[-h](\lambda, x) = \text{g-lim sup}_{\mu \downarrow 0} \partial_{(x, \lambda)}[-h]_\mu(\lambda, x)$, with the subgradients as in (10).*

Proof. **Proposition 15** states $h_\mu(\lambda, x)$ is strictly increasing in μ , and $h_\mu(\lambda, x) \geq h(\lambda, x)$ for $\mu \geq 0$. Thus, $h_\mu(\lambda, x)$ is locally equi-bounded for $0 \leq \mu \leq 1$ because for $(x', \lambda') \in \Sigma = \mathcal{B}((\lambda, x), 1)$ we have $c_0 \leq h_\mu(\lambda', x') \leq c_1$, where the $c_\mu = \min_{x, \lambda} \{h_\mu(\lambda, x) \mid (x, \lambda) \in \Sigma\}$ are finite since Σ is compact and h_0, h_1 are locally Lipschitz (see **Proposition 14**). Next, recall $\text{e-lim}_{\mu \rightarrow 0} [-h]_\mu(\lambda, x) = [-h](\lambda, x)$ by **Proposition 15**, and $[-h]_\mu(\lambda, x)$ is equi-primal-lower-nice by **Proposition 16**. The result follows since we have shown the conditions of Theorem 3.5 in [22]. \square

COROLLARY 18. *Suppose **A1**, **A2** and **R1** hold. If $\lambda \geq 0$, then we have that $\lim_{\mu \downarrow 0} \nabla_{(x, \lambda)}[-h]_\mu(\lambda, x) \in \partial_{(x, \lambda)}[-h](\lambda, x)$.*

Proof. Note $\lim_{\mu \rightarrow 0} \sigma_\mu(\lambda, x)$ exists by Corollary 5.2 of [6], and the equations defining $\nabla_{(x, \lambda)}[-h]_\mu(\lambda, x)$ are continuous in \bar{y} by (9) and **A1**. So $\lim_{\mu \rightarrow 0} \nabla_{(x, \lambda)}[-h]_\mu(\lambda, x)$ exists. But when $\mu > 0$ we have $\{\nabla_{(x, \lambda)}[-h]_\mu(\lambda, x)\} = \partial_{(x, \lambda)}[-h]_\mu(\lambda, x)$ (compare (10) from **Proposition 14** with (9) from **Corollary 13**), and so $\lim_{\mu \rightarrow 0} \nabla_{(x, \lambda)}[-h]_\mu(\lambda, x) \in \text{g-lim sup}_{\mu \downarrow 0} \partial_{(x, \lambda)}[-h]_\mu(\lambda, x)$. The result then follows by applying **Theorem 17**. \square

4. Duality-Based Reformulation. It will be more convenient to work with the approximate bilevel programming problem, which is defined as the problem

$$\begin{aligned} \min_{x, y} \quad & F(x, y) \\ \text{BLP}(\epsilon) \quad & \text{s.t. } G(x) \leq 0 \\ & y \in \epsilon\text{-arg min}_y \{f(x, y) \mid g(x, y) \leq 0\} \end{aligned}$$

where $y \in \epsilon\text{-arg min}_y \{f(x, y) \mid g(x, y) \leq 0\}$ means $f(x, y) \leq \min_y \{f(x, y) \mid g(x, y) \leq 0\} + \epsilon$ and $g(x, y) \leq \epsilon$. (Equivalently, we have that y is an ϵ -solution in the sense of [28, 29].) Note this optimization problem is equivalent to BLP when $\epsilon = 0$.

We first define our duality-based reformulation of $\text{BLP}(\epsilon)$, and then show it has a solution and is equivalent to the approximate bilevel program. Next we study constraint qualification of our reformulation and provide conditions that ensure MFCQ holds. Since the duality-based reformulation has regularization, we conclude by providing sufficient conditions that ensure convergence of stationary points of the regularized duality-based reformulation to stationary points of the limiting problem.

4.1. Definition. Our duality-based reformulation of $\text{BLP}(\epsilon)$ using the RDF is

$$\begin{aligned} \text{DBP}(\epsilon, \mu) \quad & \min_{x, y, \lambda} F(x, y) \\ & \text{s.t. } G(x) \leq 0 \\ & f(x, y) - h_\mu(\lambda, x) \leq \epsilon, \quad g(x, y) \leq \epsilon, \quad \lambda \geq 0 \end{aligned}$$

It will be helpful to separately refer to the feasible set of $\text{DBP}(\epsilon, \mu)$, which is given by

$$(11) \quad \mathcal{C}(\epsilon, \mu) = \left\{ (x, y, \lambda) : \begin{array}{l} G(x) \leq 0 \\ f(x, y) - h_\mu(\lambda, x) \leq \epsilon, \quad g(x, y) \leq \epsilon, \quad \lambda \geq 0 \end{array} \right\}$$

One useful property of the reformulation $\text{DBP}(\epsilon, \mu)$ is it is convex when x is fixed, and this result is a generalization of Proposition 6 in [3] which has a similar proof. The significance of this pointwise convexity is twofold. It induces additional regularity on $\text{DBP}(\epsilon, \mu)$. Also, it has useful implications in the design of algorithms to solve $\text{BLP}(\epsilon)$.

PROPOSITION 19. *Suppose **A1**, **A2** and **R1** hold. Then $\text{DBP}(\epsilon, \mu)$ is a convex optimization problem when x is fixed.*

Proof. Following Example 11.46 of [36], the RDF corresponds to minimizing the generalized Lagrangian given by $\ell(x, y, \lambda) = \delta_Y(y) + \mu\|y\|^2 + f(x, y) + \langle \lambda, g(x, y) \rangle - \delta_\Lambda(\lambda)$. And since $\ell(x, y, \lambda)$ is concave in λ by Proposition 11.48 of [36], we get that $h_\mu(\lambda, x)$ is concave in λ by Theorem 5.5 of [36]. The result follows by noting this implies $f(x, y) - h_\mu(\lambda, x)$ is convex in y for fixed x . \square

Note the equivalence $h(\lambda, x) = h_0(\lambda, x)$. The following result is essentially a corollary of Theorem 8, combined with Proposition 5 of [3].

PROPOSITION 20. *Suppose **A1**, **A2** and **R1** hold. Then a point y is an ϵ -solution to the lower level problem if and only if there exists λ such that*

$$(12) \quad f(x, y) - h_0(\lambda, x) \leq \epsilon, \quad g(x, y) \leq \epsilon, \quad \lambda \geq 0$$

Proof. By Proposition 5 of [3] a point x is an ϵ -solution to the lower level problem if and only if there exists λ such that the following inequalities are satisfied: $f(x, y) - \psi(\lambda, x) \leq \epsilon$, $g(x, y) \leq \epsilon$, $\lambda \geq 0$. The result holds if we can show there exists $\lambda' \geq 0$ such that $f(x, y) - \psi(\lambda', x) \leq \epsilon$ if and only if there exists $\lambda'' \geq 0$ such that $f(x, y) - h_0(\lambda'', x) \leq \epsilon$. Let $\lambda'' \in \arg \max_{\lambda} \{h_0(\lambda, x) \mid \lambda \geq 0\}$, and note $f(x, y) - h_0(\lambda'', x) \leq f(x, y) - \psi(\lambda', x)$ by Theorem 8. Similarly, let $\lambda' \in \arg \max_{\lambda} \{\psi(\lambda, x) \mid \lambda \geq 0\}$, and note $f(x, y) - \psi(\lambda', x) \leq f(x, y) - h_0(\lambda'', x)$ by Theorem 8. \square

COROLLARY 21. *Suppose **A1**, **A2** and **R1** hold. If $\mu \geq 0$ and a point y is an ϵ -solution to the lower level problem, then there exists λ such that*

$$(13) \quad f(x, y) - h_\mu(\lambda, x) \leq \epsilon, \quad g(x, y) \leq \epsilon, \quad \lambda \geq 0$$

Proof. By Proposition 20 there exists λ satisfying (12), and so $f(x, y) - h_\mu(\lambda, x) \leq f(x, y) - h_0(\lambda, x) \leq \epsilon$ since Proposition 15 shows $[-h]_\mu(\lambda, x)$ is strictly decreasing. \square

These results show that upper-bounding the objective function by the RDF is an optimality condition for the lower level problem, and this leads to an equivalence between the global minimizers of $\text{BLP}(\epsilon)$ and $\text{DBP}(\epsilon, 0)$. A similar result was shown in [16] for the KKT reformulation, but we cannot apply their results to our setting because feasible λ for $\text{DBP}(\epsilon)$ are not necessarily Lagrange multipliers when $\epsilon > 0$.

PROPOSITION 22. *Suppose **A1**, **A2** and **R1** hold. A point (\bar{x}, \bar{y}) is a global minimum of $\text{BLP}(\epsilon)$ if and only if for some feasible $\lambda \geq 0$ the point $(\bar{x}, \bar{y}, \lambda)$ is a global minimum of $\text{DBP}(\epsilon, 0)$.*

Proof. We prove this by showing a point (x', y') is not a global minimum of $\text{BLP}(\epsilon)$ if and only if (x', y', λ') is not a global minimum of $\text{DBP}(\epsilon, 0)$ for some feasible $\lambda' \geq 0$. Suppose (x', y') is not a global minimum of $\text{BLP}(\epsilon)$. Then there exists (x, y) feasible for $\text{BLP}(\epsilon)$, and such that $F(x, y) < F(x', y')$. By **Proposition 20**, there exists $\lambda \geq 0$ such that (x, y, λ) is feasible for $\text{DBP}(\epsilon, 0)$, which implies (x', y', λ') is not a global minimum of $\text{DBP}(\epsilon, 0)$. Similarly, suppose (x', y', λ') is not a global minimum of $\text{DBP}(\epsilon, 0)$. Then there exists (x, y, λ) feasible for $\text{DBP}(\epsilon, 0)$, and such that $F(x, y) < F(x', y')$. However, this (x, y) is feasible for $\text{BLP}(\epsilon)$ by **Proposition 20**. Thus (x', y') is not a global minimum of $\text{BLP}(\epsilon)$. \square

There is a potential issue with the above result. In particular, we have not shown that $\text{DBP}(\epsilon, \mu)$ has a solution (i.e., whether the minimum is finite and the set of minimizers is nonempty and compact). Though **A1–A3** and **R1, R2** ensure $\text{BLP}(\epsilon)$ has a solution (see **Corollary 24**), one could imagine some pathology caused by the λ variables that causes $\text{DBP}(\epsilon, \mu)$ to not have a solution. We prove this is not the case.

PROPOSITION 23. *Suppose **A1–A3** and **R1, R2** hold. Then $\text{DBP}(\epsilon, \mu)$ has a solution, meaning the minimum exists and the set of minimizers is nonempty and compact.*

Proof. $\mathcal{C}(\epsilon, \mu)$ is closed because of **A1, A3** and **Proposition 14**, and so the result follows if we prove $\mathcal{C}(\epsilon, \mu)$ is bounded and nonempty. Note $\{x : (x, y, \lambda) \in \mathcal{C}(\epsilon, \mu)\}$ is bounded by **R2**. Next observe $\phi_\epsilon(x) = \{y : g(x, y) \leq \epsilon\}$ is: continuous by **A1, R1** and Example 5.10 of [36]; bounded for each $x \in X$ by **A1, A2, R1** and Corollary 8.7.1 of [35], and convex for each $x \in X$ by **A1**. This means we can apply **Lemma 1**, which implies $\{y : (x, y, \lambda) \in \mathcal{C}(\epsilon, \mu)\}$ is bounded.

We next show $\{\lambda : (x, y, \lambda) \in \mathcal{C}(\epsilon, \mu)\}$ is bounded by considering different cases. The set $\Phi_{\epsilon, \mu}(x) = \{(y, \lambda) : (x, y, \lambda) \in \mathcal{C}(\epsilon, \mu)\}$ is used, and observe it is convex-valued by **Proposition 19**. The first case is $\epsilon = 0$ and $\mu = 0$. Then **Theorem 8** and **Corollary 10** imply $\Phi_{0,0}(x)$ consists of saddle points to the Lagrangian \mathcal{L} , and hence satisfy the KKT conditions (see for instance Corollary 11.51 of [36]) because of the constraint qualification from **R1**. Now supposing the set $\Phi_{0,0}$ is not bounded, there exists a sequence $(x^\nu, y^\nu, \lambda^\nu) \in \mathcal{C}(0, 0)$ such that $\|\lambda^\nu\| \rightarrow \infty$. Since x^ν, y^ν (as shown above) and $\lambda^\nu / \|\lambda^\nu\|$ (which has a norm of 1) are bounded, there is some convergent subsequence by the Bolzano-Weierstrass theorem; and so by extracting this subsequence we can assume $(x^\nu, y^\nu, \lambda^\nu / \|\lambda^\nu\|) \rightarrow (\bar{x}, \bar{y}, \underline{\lambda})$ for some finite $(\bar{x}, \bar{y}, \underline{\lambda})$. (Note $g(\bar{x}, \bar{y}) \leq \epsilon$ because this set is closed by **A1**, and $\underline{\lambda} \geq 0$ since $\lambda \geq 0$ is a cone.) But $(y^\nu, \lambda^\nu) \in \Phi_{0,0}(x^\nu)$, and so (y^ν, λ^ν) satisfy the KKT conditions – particularly stationarity and complimentary slackness: $\nabla_y f(x^\nu, y^\nu) + \nabla_y g(x^\nu, y^\nu)' \lambda^\nu = 0$ and $\lambda_i^\nu g_i(x^\nu, y^\nu) = 0$. Dividing by $\|\lambda^\nu\|$ and taking the limit, we get $\nabla_y g(\bar{x}, \bar{y})' \underline{\lambda} = 0$ and $\underline{\lambda}_i g_i(\bar{x}, \bar{y}) = 0$, where we have used that $f, g \in \mathcal{C}^2$ (i.e., **A1**). This means the MFCQ do not hold at \bar{y} for \bar{x} , which is a contradiction on **R1, R2** since Slater's condition is equivalent to MFCQ for convex sets [36]. This shows $\Phi_{0,0}$ is bounded. Next consider the case where $\epsilon > 0$. For each $x \in X$, choosing \bar{y} to be a solution to

the lower level problem (which exists by [Proposition 6](#)) gives a corresponding $\underline{\lambda}$ (by [Proposition 20](#)) that satisfies $f(x, \bar{y}) - h_0(x, \underline{\lambda}) \leq 0 < \epsilon$. But h_μ is strictly decreasing in μ ([Proposition 15](#)), and so we have $f(x, \bar{y}) - h_\mu(x, \underline{\lambda}) < \epsilon$. Since h_μ is continuous by [Proposition 14](#), this means we can choose $\underline{\lambda}' > 0$ such that $f(x, \bar{y}) - h_\mu(x, \underline{\lambda}') < \epsilon$. Combining this with **A1**, [Propositions 14](#) and [19](#), and Example 5.10 of [\[36\]](#) shows $\Phi_{\epsilon, \mu}$ is continuous. Noting $\Phi_{0,0}$ is bounded, it follows from Corollary 8.7.1 of [\[35\]](#) and [Proposition 19](#) that $\Phi_{\epsilon, \mu}$ is also bounded for each $x \in X$. Moreover, $\Phi_{\epsilon, \mu}$ is convex for each $x \in X$ by [Proposition 19](#). This means we can apply [Lemma 1](#), which implies $\Phi_{\epsilon, \mu}$ is bounded. The last case is when $\mu > 0$. For each $x \in X$, choosing \bar{y} to be a solution to the lower level problem (which exists by [Proposition 6](#)) gives a corresponding $\underline{\lambda}$ (by [Proposition 20](#)) that satisfies $f(x, \bar{y}) - h_0(x, \underline{\lambda}) \leq 0 \leq \epsilon$. But h_μ is strictly decreasing in μ ([Proposition 15](#)), and so $f(x, \bar{y}) - h_\mu(x, \underline{\lambda}) < \epsilon$. The remaining argument is identical to the second case. This finishes showing $\{\lambda : (x, y, \lambda) \in \mathcal{C}(\epsilon, \mu)\}$ is bounded.

We have shown $\mathcal{C}(x, y, \lambda)$ is bounded and closed. It is also nonempty since for any $x \in X$ (which is nonempty by **R2**) there is a solution to the lower level problem (by [Proposition 6](#)), and so there exists a y and $\lambda \geq 0$ such that $g(x, y) \leq \epsilon$ and $f(x, y) - h_\mu(\lambda, x) \leq \epsilon$ (by [Corollary 21](#)). By noting F is continuous by **A3**, the result follows from Example 1.11 of [\[36\]](#). \square

COROLLARY 24. *Suppose **A1–A3** and **R1, R2** hold. Then $\text{BLP}(\epsilon)$ has a solution, meaning the minimum exists and the set of minimizers is nonempty and compact.*

Proof. [Proposition 23](#) implies $\text{DBP}(\epsilon, 0)$ has a solution, meaning the global minimum exists and the set of minimizers $\Phi = \arg \min \text{DBP}(\epsilon, 0)$ is nonempty and compact. [Proposition 22](#) implies $\Phi' = \{(x, y) : (x, y, \lambda) \in \Phi\} = \arg \min \text{BLP}(\epsilon)$. But Φ' is the projection of a compact and nonempty set, and so Φ' is also compact and nonempty. \square

The issue of equivalence between local minimizers of $\text{BLP}(\epsilon)$ and $\text{DBP}(\epsilon, 0)$ is more complex. The KKT reformulation lacks such an equivalence when there is a generic lack of continuity of Lagrange multipliers of the lower level problem [\[16\]](#), and [\[16\]](#) argues that assuming LICQ for the lower level problem provides equivalence of local minimizers since this ensures uniqueness (and hence continuity) of the Lagrange multipliers [\[44\]](#). However, results for the KKT reformulation [\[16\]](#) cannot be applied to our setting because feasible λ for $\text{DBP}(\epsilon, 0)$ are not necessarily Lagrange multipliers.

PROPOSITION 25. *Suppose **A1, A2** and **R1** hold. If (a) $\epsilon > 0$, or (b) $g(x, y)$ satisfies LICQ for each $x \in X$; then a point (\bar{x}, \bar{y}) is a local minimum of $\text{BLP}(\epsilon)$ if and only if for some feasible $\lambda \geq 0$ the point $(\bar{x}, \bar{y}, \lambda)$ is a local minimum of $\text{DBP}(\epsilon, 0)$.*

Proof. We show (x', y') is not a local minimum of $\text{BLP}(\epsilon)$ if and only if (x', y', λ') is not a local minimum of $\text{DBP}(\epsilon, 0)$ for some feasible $\lambda' \geq 0$. First suppose (x', y', λ') is not a local minimum of $\text{DBP}(\epsilon, 0)$. Then there exists a feasible sequence $(x^\nu, y^\nu, \lambda^\nu) \rightarrow (x', y', \lambda')$ with $F(x^\nu, y^\nu) < F(x', y')$, where (x^ν, y^ν) is feasible for $\text{BLP}(\epsilon)$ by [Proposition 20](#). This shows (x', y') is not a local minimum of $\text{BLP}(\epsilon)$. To prove the other direction, suppose (x', y') is not a local minimum of $\text{BLP}(\epsilon)$. Then there exists a sequence of feasible $(x^\nu, y^\nu) \rightarrow (x', y')$ with $F(x^\nu, y^\nu) < F(x', y')$. We must consider two cases, and the first is when $\epsilon > 0$. In the proof of [Proposition 23](#), we showed $\Phi_{\epsilon, \mu}(x) = \{(y, \lambda) : (x, y, \lambda) \in \mathcal{C}(\epsilon, \mu)\}$ is continuous when $\epsilon > 0$, and so there exists a sequence $\lambda^\nu \rightarrow \lambda'$ with $(x^\nu, y^\nu, \lambda^\nu)$ feasible for $\text{DBP}(\epsilon, 0)$. This implies (x', y', λ') is not a local minimum of $\text{DBP}(\epsilon, 0)$. The second case is when $\epsilon = 0$ and LICQ holds. [Theorem 8](#) and [Corollary 10](#) imply $\Phi_{0,0}(x)$ consists of saddle points to the Lagrangian \mathcal{L} , and hence satisfy the KKT conditions (see Corollary 11.51 of [\[36\]](#)) because of the constraint qualification in **R1**. So there is a unique $\lambda'(x)$ that makes $(x', y', \lambda'(x))$

feasible for $\text{DBP}(\epsilon, 0)$ [44]. By [Corollary 10](#) we have $\lambda'(x) \in \arg \max_{\lambda} h_0(\lambda, x)$, and so $\lambda'(x)$ is a continuous function since it is single-valued [44] and osc by the Berge maximum theorem [7]. Hence there exists a sequence $\lambda^\nu \rightarrow \lambda'(x)$ with $(x^\nu, y^\nu, \lambda^\nu)$ feasible for $\text{DBP}(\epsilon, 0)$. This implies (x', y', λ') is not a local minimum of $\text{DBP}(\epsilon, 0)$. \square

4.2. Constraint Qualification. One difficulty with solving bilevel programs is that basic reformulations do not satisfy constraint qualification [39, 40, 45]. The issue is not that the feasible region of a bilevel program usually has no interior, but rather that an inequality representing optimality must fundamentally violate constraint qualification since we can interpret constraint qualification as stating the constraints have no local optima [31]. Given this intuition, the following result is not surprising:

PROPOSITION 26. *Suppose **A1–A3** and **R1** hold. Then MFCQ fails for $\text{DBP}(0, 0)$.*

Proof. By the subgradient and local Lipschitz continuity of $-h(\lambda, x)$ from [Proposition 15](#), and Exercise 10.10 of [36]; we have that the subgradient of $\omega(x, y, \lambda) = f(x, y) - h_0(\lambda, x)$ is $\partial_x \omega(x, y, \lambda) = -\text{co}\{\langle \lambda, \nabla_x g(x, \bar{y}) \rangle \mid \bar{y} \in \sigma(\lambda, x)\}$, $\partial_y \omega(x, y, \lambda) = \{\nabla_y f(x, y)\}$, and $\partial_\lambda \omega(x, y, \lambda) = -\text{co}\{g(x, \bar{y}) \mid \bar{y} \in \sigma(\lambda, x)\}$. Now suppose we choose any $x \in X$ such that there exists a y satisfying $g(x, y) \leq 0$. (If no such x exists, then the problem is infeasible and so MFCQ trivially does not hold.) Select any $y \in s(x)$, let λ be the corresponding Lagrange multipliers (which must exist by [Proposition 6](#)), and note $s(x) = \sigma(\lambda, x)$ for this choice by [Corollary 10](#). Next, note the gradient of the constraints $-\lambda \leq 0$ is a vector whose entries are -1 . For such a choice of (x, y, λ) ,

$$(14) \quad \begin{aligned} \partial_x \omega(x, y, \lambda) + \nabla_x g(x, y)' \lambda + \nabla_x G(x)' 0 &\ni 0 \\ \partial_y \omega(x, y, \lambda) + \nabla_y g(x, y)' \lambda &\ni 0 \\ \partial_\lambda \omega(x, y, \lambda) + -1'[-g](x, y) &\ni 0 \end{aligned}$$

due to the subgradient of $\omega(x, y, \lambda)$, and the second line uses KKT stationarity since $y \in s(x)$. Note the constraint $\omega(x, y, \lambda)$ is always active by strong duality (see [Corollary 10](#)). By KKT complementary slackness, $g_i(x, y) < 0$ implies $\lambda_i = 0$; this means $\nabla_x g(x, y)' \lambda = \sum_{i \in I} \lambda_i \nabla_x g_i(x, y)'$ and $\nabla_y g(x, y)' \lambda = \sum_{i \in I} \lambda_i \nabla_y g_i(x, y)'$, where I are indices of active constraints in $g(x, y)$. Also, $\nabla_x G(x)' 0 = \sum_{j \in J} 0 \cdot \nabla_x G_j(x)'$, where J are indices of active constraints in $G(x)$. So MFCQ fails since (14) shows a nonzero multiplier makes the sum of active constraint subgradients contain the zero vector. \square

However, one of the important benefits of regularization (via having either $\epsilon > 0$ or $\mu > 0$) is that it leads to constraint qualification of the regularized problem $\text{DBP}(\epsilon, \mu)$.

THEOREM 27. *Suppose **A1–A3** and **R1, R2** hold. If $\epsilon > 0$ or $\mu > 0$, then MFCQ holds for $\text{DBP}(\epsilon, \mu)$.*

Proof. Consider any (x, y, λ) feasible for $\text{DBP}(\epsilon, \mu)$. Note some subset of the constraints $g(x, y) \leq \epsilon$, $G(x) \leq 0$, and $\lambda \geq 0$ may be active, and label the indices of the active constraints by I, J, K . Note Slater's condition holds for $g(x, y) \leq \epsilon$ by **R1**, MFCQ holds for $G(x) \leq 0$ by **R2**, and Slater's condition holds for $-\lambda \leq 0$ since it clearly has an interior. Since Slater's condition is equivalent to MFCQ for convex sets [36], there exists d_x, d_y, d_λ such that $\langle \nabla_x G_i(x), d_x \rangle < 0$, $\langle \nabla_y g_j(x, y), d_y \rangle < 0$, and $\langle \nabla_\lambda [-\lambda]_k, d_\lambda \rangle < 0$ for active constraints $i \in I$, $j \in J$, and $k \in K$.

We first prove the result for the case where $\epsilon > 0$, and note there are two sub-cases. The first sub-case has $f(x, y) - h_\mu(\lambda, x) < \epsilon$, which means this constraint cannot be active. Note that we can choose $\gamma > 0$ small enough to ensure $\langle \nabla_x G_i(x), \gamma \cdot d_x \rangle < 0$ and $\langle \nabla_x g_j(x, y), \gamma \cdot d_x \rangle + \langle \nabla_y g_j(x, y), d_y \rangle < 0$ for $i \in I$ and $j \in J$, since by the Cauchy-Schwartz inequality we have $\langle \nabla_x g_j(x, y), \gamma \cdot d_x \rangle \leq \gamma \cdot \|\nabla_x g_j(x, y)\| \cdot \|d_x\|$.

Thus, MFCQ holds in this sub-case. In the second sub-case, $f(x, y) - h_\mu(\lambda, x) = \epsilon$. Let $y^* \in \arg \min\{f(x, y) \mid g(x, y) \leq 0\}$ and $\lambda^* \in \arg \max\{h(\lambda, x) \mid \lambda \geq 0\}$, and note [Corollary 10](#) gives $f(x, y^*) - h_0(\lambda^*, x) \leq 0$. This implies $f(x, y^*) - h_0(\lambda^*, x) < \epsilon$. But recall [Proposition 15](#) gives that $[-h]_\mu(\lambda, x)$ is strictly decreasing in μ , and so we have

$$(15) \quad f(x, y^*) - h_\mu(\lambda^*, x) \leq f(x, y^*) - h_0(\lambda^*, x) < \epsilon = f(x, y) - h_\mu(\lambda, x).$$

Consider any $v_x \in \partial_x[-h]_\mu(\lambda, x)$ and $v_\lambda \in \partial_\lambda[-h]_\mu(\lambda, x)$, where existence and boundedness of the subgradient comes from [Proposition 14](#). Observe that $f(x, y) - h_\mu(\lambda, x)$ is convex in y, λ by [Proposition 19](#), and by its convexity we have $\langle \nabla_y f(x, y), y^* - y \rangle + \langle v_\lambda, \lambda^* - \lambda \rangle \leq f(x, y^*) - h_\mu(\lambda^*, x) - f(x, y) + h_\mu(\lambda, x) < 0$, where we have used (15). Since the subgradient of $[-h]_\mu$ is bounded, by the Cauchy-Schwarz inequality we can choose $\gamma > 0$ small enough to ensure $\langle \nabla_x G_i(x), \gamma \cdot d_x \rangle < 0$, $\langle \nabla_x g_j(x, y), \gamma \cdot d_x \rangle + \langle \nabla_y g_j(x, y), d_y \rangle < 0$, and $\langle \nabla_x f(x, y) + v_x, \gamma \cdot d_x \rangle + \langle \nabla_y f(x, y), y^* - y \rangle + \langle v_\lambda, \lambda^* - \lambda \rangle < 0$, for $i \in I$ and $j \in J$. We next compute $\langle \nabla_\lambda[-\lambda]_k, \lambda^* - \lambda \rangle$ for $k \in K$. If the constraint $-\lambda_k \leq 0$ is active, then $\lambda_k = 0$ and $\lambda_k^* - \lambda_k \geq 0$ because $\lambda_k^* \geq 0$. Thus, $\langle \nabla_\lambda[-\lambda]_k, \lambda^* - \lambda \rangle \leq 0$ for $k \in K$. Next, note we can choose γ' such that $\langle \nabla_\lambda[-\lambda]_k, \lambda^* - \lambda \rangle + \langle \nabla_\lambda[-\lambda]_k, \gamma' \cdot d_\lambda \rangle < 0$ for $k \in K$ and $\langle \nabla_x f(x, y) + v_x, \gamma \cdot d_x \rangle + \langle \nabla_y f(x, y), y^* - y \rangle + \langle v_\lambda, \lambda^* - \lambda \rangle + \langle v_\lambda, \gamma' \cdot d_\lambda \rangle < 0$. Thus, MFCQ holds in this sub-case.

We next prove the result for the case where $\mu > 0$, and note there are two sub-cases. The first sub-case has $f(x, y) - h_\mu(\lambda, x) < \epsilon$, which means this constraint cannot be active. This sub-case is then identical to the first sub-case above, and the proof follows accordingly. The second sub-case is when $f(x, y) - h_\mu(\lambda, x) = \epsilon$. Let $y^* \in \arg \min\{f(x, y) \mid g(x, y) \leq 0\}$ and $\lambda^* \in \arg \max\{h(\lambda, x) \mid \lambda \geq 0\}$, and note that [Corollary 10](#) gives $f(x, y^*) - h_0(\lambda^*, x) \leq 0$. But recall (as shown in the proof of [Proposition 15](#)) that $[-h]_\mu(\lambda, x)$ is strictly decreasing in μ , and so we have that

$$(16) \quad f(x, y^*) - h_\mu(\lambda^*, x) < f(x, y^*) - h_0(\lambda^*, x) \leq \epsilon = f(x, y) - h_\mu(\lambda, x).$$

The proof finishes identically to the second sub-case above, with (16) replacing (15). \square

4.3. Consistency of Approximation. We show that the regularized problems $\text{DBP}(\epsilon, \mu)$ are consistent approximations [\[31, 38\]](#) of the limiting problem $\text{DBP}(\bar{\epsilon}, 0)$ under appropriate conditions. Our first result concerns convergence of the constraint sets $\mathcal{C}(\epsilon, \mu)$, which leads as a corollary to convergence of global optimizers of the regularized problems to global optimizers of the limiting problem.

PROPOSITION 28. *Suppose **A1–A3** and **R1** hold. Then for any $\bar{\epsilon} \geq 0$ we have that $\lim_{\epsilon \downarrow \bar{\epsilon}, \mu \downarrow 0} \mathcal{C}(\epsilon, \mu) = \mathcal{C}(\bar{\epsilon}, 0)$, and $\mathcal{C}(\epsilon_1, \mu_1) \supseteq \mathcal{C}(\epsilon_2, \mu_2)$ whenever $\epsilon_1 \geq \epsilon_2$ and $\mu_1 \geq \mu_2$.*

Proof. For any $(x, y, \lambda) \in \mathcal{C}(\epsilon_2, \mu_2)$, we have: $G(x) \leq 0$, $f(x, y) - h_{\mu_2}(\lambda, x) \leq \epsilon_2$, $g(x, y) \leq \epsilon_2$, and $\lambda \geq 0$. [Proposition 15](#) shows $[-h]_\mu(\lambda, x)$ is strictly decreasing in μ , and so $f(x, y) - h_{\mu_1}(\lambda, x) \leq f(x, y) - h_{\mu_2}(\lambda, x) \leq \epsilon_2 \leq \epsilon_1$. Similarly, $g(x, y) \leq \epsilon_1 \leq \epsilon_2$. This shows $(x, y, \lambda) \in \mathcal{C}(\epsilon_1, \mu_1)$, which proves $\mathcal{C}(\epsilon_1, \mu_1) \supseteq \mathcal{C}(\epsilon_2, \mu_2)$ whenever $\epsilon_1 \geq \epsilon_2$ and $\mu_1 \geq \mu_2$. But $\mathcal{C}(\epsilon, \mu)$ is closed since f, g, G are differentiable by **A1, A3**; and h_μ is continuous by [Proposition 14](#). So the result follows by Exercise 4.3.b of [\[36\]](#). \square

COROLLARY 29. *Suppose **A1–A3** and **R1, R2** hold. If $\epsilon \downarrow \bar{\epsilon}, \mu \downarrow 0, z \downarrow 0$, then $\limsup_{\epsilon \downarrow \bar{\epsilon}, \mu \downarrow 0, z \downarrow 0} (z - \arg \min \text{DBP}(\epsilon, \mu)) \subseteq \arg \min \text{DBP}(\bar{\epsilon}, 0)$ and $z - \min \text{DBP}(\epsilon, \mu) \rightarrow \min \text{DBP}(\bar{\epsilon}, 0)$.*

Proof. Note $\text{DBP}(\epsilon, \mu)$ is $\tilde{f}_{\epsilon, \mu}(x, y, \lambda) = f(x, y) + \delta_{\mathcal{C}(\epsilon, \mu)}(x, y, \lambda)$. Recall $\mathcal{C}(\epsilon, \mu)$ is closed since: f, g, G are differentiable by **A1, A3**; and h_μ is continuous by [Proposition 14](#). By [Proposition 28](#) we have $\mathcal{C}(\epsilon_1, \mu_1) \supseteq \mathcal{C}(\epsilon_2, \mu_2)$ when $\epsilon_1 \geq \epsilon_2$ and $\mu_1 \geq \mu_2$,

and so $\tilde{f}_{\epsilon_1, \mu_1} \leq \tilde{f}_{\epsilon_2, \mu_2}$ for $\epsilon_1 \geq \epsilon_2$ and $\mu_1 \geq \mu_2$. This means by Proposition 7.4.d of [36] that $\text{DBP}(\epsilon, \mu)$ epi-converges to $\text{DBP}(\bar{\epsilon}, 0)$ as $\epsilon \downarrow \bar{\epsilon}, \mu \downarrow 0$. Moreover, $\mathcal{C}(0, 0)$ is nonempty by **R1** and Proposition 20, which implies $\mathcal{C}(\epsilon, \mu)$ is nonempty. So $\tilde{f}_{\epsilon, \mu}(x, y, \lambda)$ is lower semicontinuous and feasible. But $\tilde{f}_{\epsilon, \mu}(x, y, \lambda) \geq \tilde{f}_{\bar{\epsilon}+1, 1}(x, y, \lambda)$ for $\epsilon \leq \bar{\epsilon}+1$ and $\mu \leq 1$ by Proposition 28. Moreover, the set $\mathcal{C}(\bar{\epsilon}+1, 1)$ is compact (as shown in the proof of Proposition 23). This means $\tilde{f}_{\epsilon, \mu}(x, y, \lambda)$ is level bounded (see Definition 1.8 in [36]) for all $\epsilon \leq \bar{\epsilon}+1$ and $\mu \leq 1$. The result now follows by Theorem 7.33 of [36]. \square

To study convergence of stationary points of the regularized problems $\text{DBP}(\epsilon, \mu)$ to stationary points of the limiting problem $\text{DBP}(\bar{\epsilon}, 0)$, we need two definitions. A stationary point (x, y, λ) of $\text{DBP}(\epsilon, \mu)$ satisfies $\nabla f(x, y) + \hat{N}_{\mathcal{C}(\epsilon, \mu)}(x, y, \lambda) \ni 0$. This is the most basic necessary condition for optimality [36]. We also define a new stationarity concept: A U-stationary point (x, y, λ) of $\text{DBP}(\bar{\epsilon}, 0)$ satisfies $\nabla f(x, y) + [\text{g-lim sup}_{\epsilon \downarrow \bar{\epsilon}, \mu \downarrow 0} \hat{N}_{\mathcal{C}(\epsilon, \mu)}](x, y, \lambda) \ni 0$. Note $\hat{N}_{\mathcal{C}(\epsilon, \mu)} = N_{\mathcal{C}(\epsilon, \mu)}$ by Proposition 36. The reason for our definition is that U-stationarity is necessary for optimality:

PROPOSITION 30. *Suppose **A1–A3** and **R1** hold. If (x, y, λ) is a local minimum of $\text{DBP}(\bar{\epsilon}, 0)$, then (x, y, λ) is a U-stationary point of $\text{DBP}(\bar{\epsilon}, 0)$.*

Proof. Exercise 6.18 of [36] and Proposition 28 imply that $\hat{N}_{\mathcal{C}(\bar{\epsilon}, 0)}(x, y, \lambda) \subseteq [\text{g-lim sup}_{\epsilon \downarrow \bar{\epsilon}, \mu \downarrow 0} \hat{N}_{\mathcal{C}(\epsilon, \mu)}](x, y, \lambda)$. And so $\nabla f(x, y) + \hat{N}_{\mathcal{C}(\bar{\epsilon}, 0)}(x, y, \lambda) \subseteq \nabla f(x, y) + [\text{g-lim sup}_{\epsilon \downarrow \bar{\epsilon}, \mu \downarrow 0} \hat{N}_{\mathcal{C}(\epsilon, \mu)}](x, y, \lambda)$. The result follows by noting Theorem 6.12 of [36] states $\nabla f(x, y) + \hat{N}_{\mathcal{C}(\bar{\epsilon}, 0)}(x, y, \lambda) \ni 0$ whenever (x, y, λ) is a local minimum, which implies $\nabla f(x, y) + [\text{g-lim sup}_{\epsilon \downarrow \bar{\epsilon}, \mu \downarrow 0} \hat{N}_{\mathcal{C}(\epsilon, \mu)}](x, y, \lambda) \ni 0$. \square

Since $\text{DBP}(\epsilon, \mu)$ is used as an approximation of solving $\text{DBP}(\bar{\epsilon}, 0)$, it is relevant to study if cluster points of stationary points to $\text{DBP}(\epsilon, \mu)$ are stationary points of $\text{DBP}(\bar{\epsilon}, 0)$. More generally, we are interested in cluster points of approximate stationary points. A ζ -approximate stationary point (x, y, λ) of $\text{DBP}(\epsilon, \mu)$ satisfies $\nabla f(x, y) + \hat{N}_{\mathcal{C}(\epsilon, \mu)}(x, y, \lambda) \ni \zeta$. Note $\hat{N}_{\mathcal{C}(\epsilon, \mu)} = N_{\mathcal{C}(\epsilon, \mu)}$ because of Proposition 36. The next result shows cluster points of stationary points are U-stationary points.

COROLLARY 31. *Suppose **A1–A3** and **R1** hold. If (x, y, λ) are ζ -approximate stationary for $\text{DBP}(\epsilon, \mu)$, then any cluster point $(\bar{x}, \bar{y}, \bar{\lambda}) \in \limsup_{\epsilon \downarrow \bar{\epsilon}, \mu \downarrow 0, \zeta \downarrow 0} (x, y, \lambda)$ is U-stationary for $\text{DBP}(\bar{\epsilon}, 0)$.*

Proof. It follows by Theorem 5.37.a of [36] and the U-stationary point definition. \square

A stronger result is when cluster points of stationary points to $\text{DBP}(\epsilon, \mu)$ are stationary points of $\text{DBP}(\bar{\epsilon}, 0)$. However, this generally requires additional regularity.

COROLLARY 32. *Suppose **A1–A3** and **R1** hold. If $\mathcal{C}(\epsilon, \mu)$ is equi-prox-regular for $\epsilon \geq \bar{\epsilon}$ and $\mu \geq 0$, then: 1. Any $(\bar{x}, \bar{y}, \bar{\lambda}) \in \limsup_{\epsilon \downarrow \bar{\epsilon}, \mu \downarrow 0, \zeta \downarrow 0} (x, y, \lambda)$ is stationary for $\text{DBP}(\bar{\epsilon}, 0)$, where the points (x, y, λ) are ζ -approximate stationary for $\text{DBP}(\epsilon, \mu)$; and 2. Any stationary point $(\bar{x}, \bar{y}, \bar{\lambda})$ of $\text{DBP}(\bar{\epsilon}, 0)$, has $\zeta(\epsilon, \mu)$ with $\lim_{\epsilon \downarrow \bar{\epsilon}, \mu \downarrow 0, \zeta \downarrow 0} (x, y, \lambda) = (\bar{x}, \bar{y}, \bar{\lambda})$, where (x, y, λ) are $\zeta(\epsilon, \mu)$ -approximate stationary for $\text{DBP}(\epsilon, \mu)$.*

Proof. Combining Corollary 3.2 of [46] with Proposition 28, we have $\hat{N}_{\mathcal{C}(\bar{\epsilon}, 0)} = \text{g-lim}_{\epsilon \downarrow \bar{\epsilon}, \mu \downarrow 0} \hat{N}_{\mathcal{C}(\epsilon, \mu)}$. The result then follows from Theorem 5.37 of [36]. \square

The above result gives conditions under which: cluster points of stationary points of the approximating problems $\text{DBP}(\epsilon, \mu)$ are stationary points of the limiting problem $\text{DBP}(\bar{\epsilon}, 0)$, and all stationary points of the limiting problem $\text{DBP}(\bar{\epsilon}, 0)$ are approachable by stationary points of the approximating problems $\text{DBP}(\epsilon, \mu)$. The sufficient

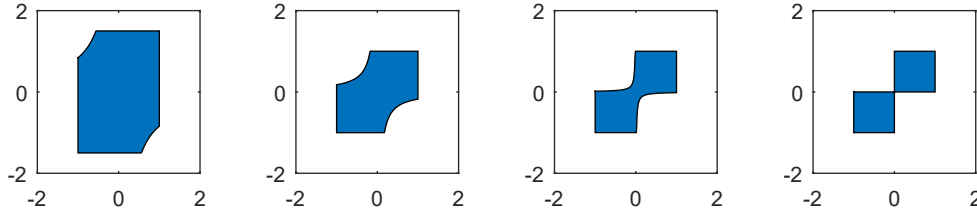
condition needed is equi-prox-regularity of the constraint sets $\mathcal{C}(\epsilon, \mu)$. It is the case that for fixed $\epsilon > 0$ or $\mu > 0$, the set $\mathcal{C}(\epsilon, \mu)$ is prox-regular.

PROPOSITION 33. *Suppose **A1–A3** and **R1, R2** hold. If $\epsilon > 0$ or $\mu > 0$, then the constraint set $\mathcal{C}(\epsilon, \mu)$ is prox-regular.*

Proof. Consider $F'(x, y, \lambda) = (G(x), f(x, y) - h_\mu(\lambda, x) - \epsilon, g(x, y) - \epsilon, -\lambda)$ and $G'(x, y, \lambda, y') = (G(x), f(x, y) - \mu\|y'\|^2 - f(x, y') - \langle \lambda, g(x, y') \rangle - \epsilon, g(x, y) - \epsilon, -\lambda)$. Note $F'(x, y, \lambda) = \max_{y' \in Y} \{G'(x, y, \lambda, y') \mid y' \in Y\}$ and $\mathcal{C}(\epsilon, \mu) = \{(x, y, \lambda) : F'(x, y, \lambda) \leq 0\}$. Since MFCQ holds by [Theorem 27](#), the result follows from [Lemma 4](#). \square

Given this result on prox-regularity of $\mathcal{C}(\epsilon, \mu)$ when $\epsilon > 0$ or $\mu > 0$, it is tempting to expect $\mathcal{C}(\epsilon, \mu)$ is also equi-prox-regular. Unfortunately, this is not the case.

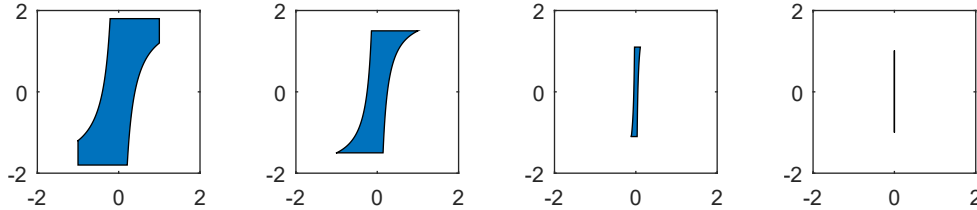
Example 34. Consider a BLP with $G(x) = (-x-1, x-1)$, $g(x, y) = (-y-1, y-1)$, $f(x, y) = (xy)^4$ if $xy \geq 0$, $f(x, y) = 0$ if $xy < 0$, and any $F \in \mathcal{C}^2$. If $Y = [-2, 2]$, then these choices satisfy **A1–A3** and **R1, R2**. When $\lambda = 0$, the constraint set is $\{(x, y, \lambda) \in \mathcal{C}(\epsilon, \mu) : \lambda = 0\} = \{(x, y) : f(x, y) \leq \epsilon, x \in [-1, 1], y \in [-1 - \epsilon, 1 + \epsilon]\}$. This set (for decreasing ϵ) is shown below, and the horizontal-vertical axes are x - y .



The final plot corresponds to $\epsilon = 0$, and we can visually confirm (see [\[1\]](#)) this constraint set is not prox-regular. Thus, the entire family of constraint sets cannot be equi-prox-regular because otherwise this would lead to a contradiction [\[46\]](#). \diamond

The issue in [Example 34](#) is not the lack of continuity of minimizers of the lower level problem with respect to x , but rather the shape of the solution mapping. The next example is a situation where the minimizers of the lower level problem are not continuous with respect to x , while the family of constraint sets is equi-prox-regular.

Example 35. Consider a BLP with $G(x) = (-x-1, x-1)$, $g(x, y) = (-y-1, y-1)$, $f(x, y) = -xy$, and any $F \in \mathcal{C}$. If $Y = [-2, 2]$, then these satisfy **A1–A3** and **R1, R2**. Note $\mathcal{C}(\epsilon, 0) = \{(x, y, \lambda) : -xy + 2|x + \lambda_1 - \lambda_2| + \lambda_1 + \lambda_2 \leq \epsilon, x \in [-1, 1], y \in [-1 - \epsilon, 1 + \epsilon], \lambda \geq 0\}$. When $\lambda = 0$, this constraint is $\{(x, y, \lambda) \in \mathcal{C}(\epsilon, 0) : \lambda = 0\} = \{(x, y) : -xy + 2|x| \leq \epsilon, x \in [-1, 1], y \in [-1 - \epsilon, 1 + \epsilon]\}$. This set (for decreasing ϵ) is shown below, and the final plot is for $\epsilon = 0$. The horizontal-vertical axes are x - y .



The entire family (i.e., not just those plotted) of $\mathcal{C}(\epsilon, 0)$ is equi-prox-regular. \diamond

The above examples show equi-prox-regularity is connected (in a nonobvious way) to the relationship between x and the solution sets of the lower level problem. Equi-prox-regularity is immediate when $\mathcal{C}(\epsilon, \mu)$ is convex, but such constraint sets correspond to easy-to-solve instances of BLP. A partial characterization of more general sufficient conditions that imply equi-prox-regularity of the sets $\mathcal{C}(\epsilon, \mu)$ is given below.

PROPOSITION 36. *Suppose **A1–A3** and **R1, R2** hold. If $\bar{\epsilon} > 0$, then the family of constraint sets $\mathcal{C}(\epsilon, \mu)$ for $\epsilon \in [\bar{\epsilon}, e]$ and $\mu \in [0, m]$ (where $e \geq \bar{\epsilon}$ and $m \geq 0$) is equi-prox-regular.*

Proof. Consider $F'(x, y, \lambda, \epsilon, \mu) = (G(x), f(x, y) - h_\mu(\lambda, x) - \epsilon, g(x, y) - \epsilon, -\lambda)$, $G'(x, y, \lambda, y') = (G(x), f(x, y) - f(x, y') - \langle \lambda, g(x, y') \rangle, g(x, y), -\lambda)$, and $H'(y', \epsilon, \mu) = (0, -\mu\|y'\|^2 - \epsilon, -\epsilon, 0)$. Note $F'(x, y, \lambda) = \max_{y' \in Y} \{G'(x, y, \lambda, y') + H'(y', \epsilon, \mu) \mid y' \in Y\}$ and $\mathcal{C}(\epsilon, \mu) = \{(x, y, \lambda) : F'(x, y, \lambda) \leq 0\}$. Since MFCQ holds by [Theorem 27](#), the result follows from [Lemma 5](#). \square

This result says $\mathcal{C}(\epsilon, \mu)$ with ϵ strictly greater than 0 form a family of constraint sets that is equi-prox-regular. This characterization is broad because it only requires mild assumptions on BLP; however, the proof of the above result does not extend to the situation where $\bar{\epsilon} = 0$, because MFCQ does not hold when $\epsilon = 0$. Characterizing sufficient conditions for equi-prox-regularity when $\bar{\epsilon} = 0$ requires further study.

5. Stability and Continuity of Solutions. We argue that approximate bilevel programs (i.e., BLP(ϵ) with $\epsilon > 0$) are better than bilevel programs BLP(0) from a modeling standpoint. We have already shown (see [Theorem 27](#)) that approximate bilevel programs satisfy MFCQ, while bilevel programs do not (see [Proposition 26](#)). [Proposition 25](#) implies local solutions of approximate bilevel programs correspond to local solutions of DBP(ϵ), which is not always true for bilevel programs. Moreover, the constraint sets of approximate bilevel programs are always guaranteed to form an equi-prox-regular family (subject to the conditions of [Proposition 36](#)). Here, we show approximate bilevel programs also have improved continuity and stability of solutions.

The value function $V_{\epsilon, \mu}(x)$ of DBP(ϵ, μ) when x is fixed is useful for studying continuity of DBP(ϵ, μ) because $\min_x V_{\epsilon, \mu}(x) = \min \text{DBP}(\epsilon, \mu)$ and $\arg \min_x V_{\epsilon, \mu}(x) = \{x : (x, y, \lambda) \in \arg \min \text{DBP}(\epsilon, \mu)\}$. It is also useful for studying continuity of BLP(ϵ) because by [Proposition 22](#) we have $\min_x V_{\epsilon, 0}(x) = \min \text{BLP}(\epsilon) = \min \text{DBP}(\epsilon, 0)$ and $\arg \min_x V_{\epsilon, 0}(x) = \{x : (x, y) \in \arg \min \text{BLP}(\epsilon)\} = \{x : (x, y, \lambda) \in \arg \min \text{DBP}(\epsilon, 0)\}$. We first generalize Proposition 7 of [\[3\]](#) for when $\epsilon > 0$, $\mu > 0$, or f is strictly convex.

PROPOSITION 37. *Suppose **A1–A3** and **R1, R2** hold. If $\epsilon > 0$, $\mu > 0$, or $f(x, y)$ is strictly convex in y for every $x \in X$, then $V_{\epsilon, \mu}(x)$ is continuous on its domain.*

Proof. We consider two cases: (a) $\epsilon > 0$ or $\mu > 0$, and (b) $\epsilon = 0$, $\mu = 0$, and f is strictly convex. When $\epsilon > 0$ or $\mu > 0$, consider $\Phi_{\epsilon, \mu}(x) = \{(y, \lambda) : (x, y, \lambda) \in \mathcal{C}(\epsilon, \mu)\}$. In the proof of [Proposition 22](#) we showed $\Phi_{\epsilon, \mu}(x)$ is continuous. This means $V_{\epsilon, \mu}$ is continuous by **A3, R2** and the Berge maximum theorem [\[7\]](#). When $\epsilon = 0$, $\mu = 0$, and f is strictly convex; note [Proposition 22](#) implies DBP(0, 0) is equivalent to BLP(0), which is easier to study. The Berge maximum theorem [\[7\]](#) gives that $s(x)$ is osc. But f is strictly convex, and so $s(x)$ is single-valued by Theorem 2.6 of [\[36\]](#). But a single-valued osc function is continuous, and so $s(x)$ is continuous in this case. Recalling we also have **A3, R2**, this means $V_{0, 0}$ is continuous by the Berge maximum theorem [\[7\]](#). \square

It is also interesting to study how DBP responds as the defining functions smoothly change. If we partition the decision variable $x = (\bar{x}, \theta)$, then we can model how

DBP(ϵ, μ) changes using the following parametric version of DBP:

$$\begin{aligned} & \min_{\bar{x}, y, \lambda} F(\bar{x}, y, \theta) \\ \text{P-DBP}(\epsilon, \mu, \theta) \quad & \text{s.t. } G(\bar{x}, \theta) \leq 0 \\ & f(\bar{x}, y, \theta) - h_\mu(\lambda, \bar{x}, \theta) \leq \epsilon, \quad g(\bar{x}, y, \theta) \leq \epsilon, \quad \lambda \geq 0 \end{aligned}$$

Let $V_{\epsilon, \mu}(\theta)$ be the value function, and $S_{\epsilon, \mu}(\theta) = \{\bar{x} : (\bar{x}, y, \lambda) \in \arg \min \text{P-BLP}(\epsilon, \mu, \theta)\}$ be the set of minimizing arguments. The following result describes their continuity.

COROLLARY 38. *Suppose **A1–A3** and **R1, R2** hold with $x = (\bar{x}, \theta)$. If $\epsilon > 0$, $\mu > 0$, or $f(\bar{x}, y, \theta)$ is strictly convex in y for every $(\bar{x}, \theta) \in X$, then $V_{\epsilon, \mu}(\theta)$ is continuous and $S_{\epsilon, \mu}(\theta)$ is osc on their domain $\text{dom}(V_{\epsilon, \mu}) = \text{dom}(S_{\epsilon, \mu}) = \{\theta : \exists \bar{x} \text{ s.t. } G(\bar{x}, \theta) \leq 0\}$.*

Proof. **Proposition 37** shows $V_{\epsilon, \mu}(\bar{x}, \theta)$ is continuous on $(\bar{x}, \theta) \in X$. But X is compact by **R2**, and so the Berge maximum theorem [7] implies $V_{\epsilon, \mu}(\theta) = \min_{\bar{x}} V_{\epsilon, \mu}(\bar{x}, \theta)$ is continuous and $S_{\epsilon, \mu}(\theta) = \arg \min_{\bar{x}} V_{\epsilon, \mu}(\bar{x}, \theta)$ is osc on their domain. \square

This result says the minimum (solutions) of approximate bilevel programs $\text{BLP}(\epsilon)$ react in a continuous (osc) manner as the parameters of the bilevel program are smoothly changed. It also says the minimum (solutions) of a bilevel program $\text{BLP}(0)$ are continuous (osc) when the objective of the lower level problem is strictly convex.

6. Numerical Examples. We conclude with two examples to demonstrate how our proposed duality-based approach can be used to solve practical bilevel programs with a convex lower level. The first is a problem of inverse optimization with noisy data [3, 8, 20], and the second involves computing a Stackelberg strategy for static routing games [4, 9, 21, 41, 43]. For exposition, we use different notation than in the above cited papers, and we consider small instances of these two problems. One important note is that since the lower level problems are convex, by choosing any $x \in X$ we can compute an initial feasible point for $\text{DBP}(\epsilon, \mu)$ by solving a convex optimization problem and keeping the optimal solution and its corresponding Lagrange multipliers.

6.1. Inverse Optimization with Noisy Data. Suppose a strategic agent makes decisions y_i in response to an external signal u_i by maximizing a parametrized utility function $U(y, u, x)$, where x is a parameter vector. One problem is to estimate x given n data points (u_i, z_i) for $i = 1, \dots, n$, where z_i are noisy measurements of y_i . Statistically consistent estimation requires solving bilevel programs [3], while heuristics using convex single-level programming (such as [8, 20]) are inconsistent [3].

Consider the instance where the agent's utility function is $U(y, u, x) = -(x + u) \cdot y$ with $x, y, u \in \mathbb{R}$. The bilevel program for the statistical estimation problem is

$$\begin{aligned} (17) \quad & \min_{x, y_i} \frac{1}{n} \sum_{i=1}^n \|z_i - y_i\|^2 \\ & \text{s.t. } x \in [-1, 1], \quad y_i \in \arg \min_y \{(x + u_i) \cdot y \mid y \in [-1, 1]\}, \forall i = 1, \dots, n \end{aligned}$$

The duality-based reformulation $\text{DBP}(\epsilon, \mu)$ for this instance is given by

$$\begin{aligned} (18) \quad & \min_{x, y_i} \frac{1}{n} \sum_{i=1}^n \|z_i - y_i\|^2 \\ & \text{s.t. } x \in [-1, 1] \\ & (x + u_i)y_i - h_\mu(\lambda_i, x) - \epsilon \leq 0, \quad y_i \in [-1 - \epsilon, 1 + \epsilon], \quad \lambda_i \geq 0, \quad \forall i = 1, \dots, n \end{aligned}$$

where $\lambda_i \in \mathbb{R}^2$, and the RDF is given by $h_\mu(\lambda_i, x) = \min_y \{\mu \cdot y^2 + (x + u_i) \cdot y + \lambda_{i,1} \cdot (-y - 1) + \lambda_{i,2} \cdot (y - 1) \mid y \in [-2, 2]\}$. Two hundred instances of this problem were

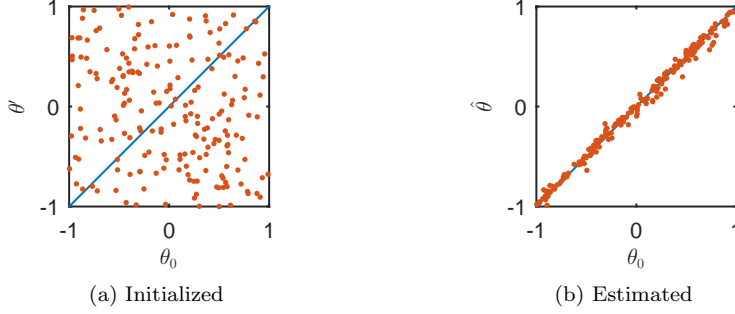


FIG. 1. Scatter plot of initialized parameters x' versus true parameter x_0 (Left). Scatter plot of estimated parameters \hat{x} versus true parameters x_0 , as computed by duality-based approach (Right).

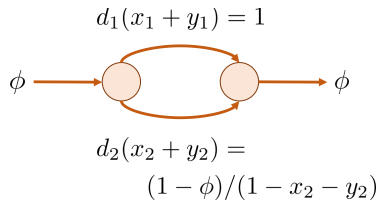
solved with $n = 100$ data points, where (a) u_i were drawn from a uniform distribution over $[-1, 1]$, (b) the true parameter x_0 was drawn from a uniform distribution over $[-1, 1]$, and (c) $z_i = \xi_i + w_i$ with $\xi_i \in \arg \min_y \{(x_0 + u_i) \cdot y \mid y \in [-1, 1]\}$ and w_i drawn from a standard normal distribution. Each instance was solved by the algorithm:

1. Initialize x' by a random draw from a uniform distribution over $[-1, 1]$.
2. For $i = 1, \dots, n$
 - (a) Choose $y'_i \in \arg \min \{(x' + u_i) \cdot y \mid y \in [-1, 1]\}$
 - (b) Set λ'_i be the corresponding Lagrange multipliers
3. Use a nonlinear solver to numerically solve DBP($10^0, 10^{-4}$): Use (x', y'_i, λ'_i) as an initial (feasible) point, and set $(x^1, y_i^1, \lambda_i^1)$ to be the computed solution
4. Use a nonlinear solver to numerically solve DBP($10^{-1}, 10^{-4}$): Use $(x^1, y_i^1, \lambda_i^1)$ as an initial point, and set $(x^2, y_i^2, \lambda_i^2)$ to be the computed solution
5. Use a nonlinear solver to numerically solve DBP($10^{-2}, 10^{-4}$): Use $(x^2, y_i^2, \lambda_i^2)$ as an initial point, and set $(\hat{x}, y_i^3, \lambda_i^3)$ to be the computed solution

Scatter plots with the two hundred instances are seen in Figure 1. The initial x' were randomly chosen, and are uncorrelated to the true parameters x_0 . Since the estimates \hat{x} were computed using the duality-based approach to solve the inverse optimization problem, there is a close correlation between the estimated and true parameters.

6.2. Stackelberg Routing Games. A common class of static routing games consists of a directed graph with multiple edges between vertices, convex delay functions for each edge, and a listing of inflows and outflows of traffic [4, 9, 21, 41, 43]. The Stackelberg strategy is a situation where a leader controls an α fraction of the flow, the remaining flow is routed according to a Nash equilibrium given the flow of the leader, and the leader routes their flow to minimize the average delay in the network. This problem can be formulated as a bilevel program with a convex lower level.

An example of a two edge network in this Stackelberg setting is shown below:



The Stackelberg strategy for this two edge network is the solution to

$$(19) \quad \begin{aligned} & \min_{x,y} x_1 + y_1 + (1 - \phi) \cdot (x_2 + y_2) / (1 - x_2 - y_2) \\ & \text{s.t. } x_1 + x_2 = \alpha \cdot \phi, x \geq 0, y \in \arg \min_y \{x_1 + y_1 + \\ & \quad - (1 - \phi) \cdot \log(1 - x_2 - y_2) \mid y_1 + y_2 = (1 - \alpha) \cdot \phi, y \geq 0\} \end{aligned}$$

where (a) x_1, x_2 is the leader's flow on the top/bottom edge, (b) y_1, y_2 is the follower's flow on the top/bottom edge, (c) $\phi < 1$ is the amount of flow entering the network, and (d) α is the fraction of the flow controlled by the leader. The duality-based reformulation $\text{DBP}(\epsilon, \mu)$ for this instance is

$$(20) \quad \begin{aligned} & \min_{x,y} x_1 + y_1 + (1 - \phi) \cdot (x_2 + y_2) / (1 - x_2 - y_2) \\ & \text{s.t. } x_1 + x_2 = \alpha \cdot \phi, \quad x, y, \lambda \geq 0 \\ & \quad x_1 + y_1 - (1 - \phi) \cdot \log(1 - x_2 - y_2) - h_\mu(\lambda, \nu, x) - \epsilon \leq 0 \\ & \quad y_1 + y_2 \in [(1 - \alpha) \cdot \phi - \epsilon, (1 - \alpha) \cdot \phi + \epsilon] \end{aligned}$$

where $\lambda \in \mathbb{R}^2$, $\nu \in \mathbb{R}$, and the RDF is $h_\mu(\lambda, \nu, x) = \min_y \{\mu \cdot \|y\|^2 + x_1 + y_1 - (1 - \phi) \cdot \log(1 - x_2 - y_2) - \lambda_1 \cdot y_1 - \lambda_2 \cdot y_2 + \nu \cdot (y_1 + y_2 - (1 - \alpha) \cdot \phi) \mid y \in [-1, 2]\}$. Different instances (corresponding to different values of α, ϕ) were solved by the following algorithm:

1. Choose $x' \in \arg \min_x \{x_1 + (1 - \phi) \cdot (x_2) / (1 - x_2) \mid x_1 + x_2 = \phi, x \geq 0\}$
2. Choose $y' \in \arg \min_y \{\alpha \cdot x'_1 + y_1 - (1 - \phi) \cdot \log(1 - \alpha \cdot x'_2 - y_2) \mid y_1 + y_2 = (1 - \alpha) \cdot \phi, y \geq 0\}$
3. Set λ', ν' to be the respective Lagrange multipliers for the inequality/equality constraints
4. Use a nonlinear solver to numerically solve $\text{DBP}(10^{-6}, 10^{-6})$: Use the point $(\alpha x', y', \lambda', \nu')$ as an initial (feasible) point, and set (x, y, λ, ν) to be the computed solution

Solution quality is evaluated by the price of anarchy (PoA) [37], which is the average delay of the solution (either $(\alpha x', y')$ for the SCALE strategy [4, 9, 43] or (x, y) for the duality-based strategy) divided by the average delay when $\alpha = 1$. The objective in (19) gives the average delay. A PoA close to 1 is ideal because it implies the delay of the strategy is close to the delay when the leader controls the entire flow, while a large PoA means the average delay of the strategy is much higher than when the leader controls the entire flow. The results in Figure 2 show that our duality-based approach (initialized with SCALE) significantly improves the quality of the Stackelberg strategy.

7. Conclusion. We defined a new (differentiable) dual function, and used this to construct a duality-based reformulation of bilevel programs with a convex lower level. This reformulation incorporates regularization to ensure constraint qualification and differentiability. We showed that when the constraints of the regularized problems are equi-prox-regular, then stationary points of the regularized problems converge to stationary points of the unregularized problem; and a sufficient condition was given for this equi-prox-regularity. Stability and continuity of solutions to approximate bilevel programs was studied, and we concluded with two numerical examples that show applicability of the duality-based reformulation to solving practical instances of BLP. One potential extension is to develop an iterative algorithm that decreases ϵ, μ at each step, as opposed to solving $\text{DBP}(\epsilon, \mu)$ to local optimality for each value of ϵ, μ (as was done in the inverse optimization with noisy data example).

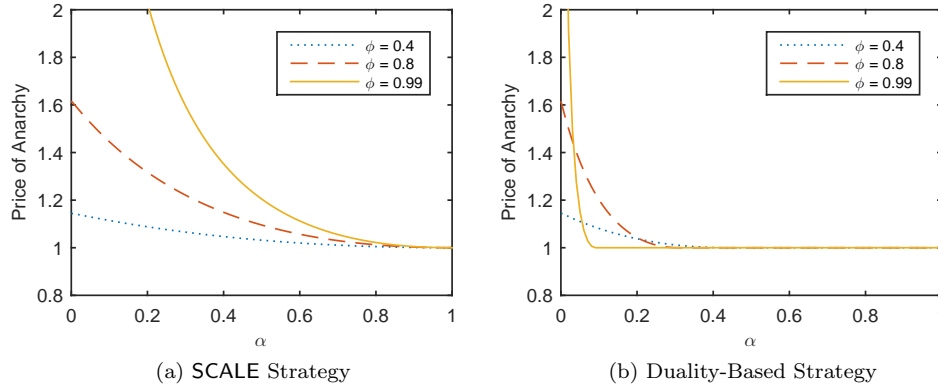


FIG. 2. Comparison of Stackelberg strategy quality for SCALE strategy (Left), which was used as an initialization to compute the duality-based strategy (Right) using our reformulation.

REFERENCES

- [1] S. ADLY, F. NACRY, AND L. THIBAUT, *Preservation of prox-regularity of sets with applications to constrained optimization*, SIAM J. Optim., 26 (2016), pp. 448–473.
- [2] M. ANITESCU, *On using the elastic mode in nonlinear programming approaches to mathematical programs with complementarity constraints*, SIAM J Optim., 15 (2005), pp. 1203–1236.
- [3] A. ASWANI, Z.-J. M. SHEN, AND A. SIDDIQ, *Inverse optimization with noisy data*, arXiv:1507.03266, (2015).
- [4] A. ASWANI AND C. TOMLIN, *Game-theoretic routing of GPS-assisted vehicles for energy efficiency*, in Proceedings of the American Control Conference, IEEE, 2011, pp. 3375–3380.
- [5] H. ATTOUCH, *Convergence de fonctions convexes, des sous-différentiels et semi-groupes associés*, CR Acad. Sci. Paris, 284 (1977), pp. 539–542.
- [6] H. ATTOUCH, *Viscosity solutions of minimization problems*, SIAM J. Optim., 6 (1996), pp. 769–806.
- [7] C. BERGE, *Topological Spaces: including a treatment of multi-valued functions, vector spaces, and convexity*, Courier Dover Publications, 1963.
- [8] D. BERTSIMAS, V. GUPTA, AND I. C. PASCHALIDIS, *Data-driven estimation in equilibrium using inverse optimization*, Mathematical Programming Series A, (2014).
- [9] V. BONIFACI, T. HARKS, AND G. SCHÄFER, *Stackelberg routing in arbitrary networks*, Mathematics of Operations Research, 35 (2010), pp. 330–346.
- [10] J. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer, 2000.
- [11] S. K. BORALUGODA, *Prox-regular functions in Hilbert spaces*, PhD thesis, University of Alberta, 1998.
- [12] J. BURKE AND T. HOHEISEL, *Epi-convergent smoothing with applications to convex composite functions*, SIAM J. Optim., 23 (2013), pp. 1457–1479.
- [13] X. CHEN, *Smoothing methods for nonsmooth, nonconvex minimization*, Mathematical Programming, 134 (2012), pp. 71–99.
- [14] F. H. CLARKE, *Generalized gradients and applications*, Transactions of the American Mathematical Society, 205 (1975), pp. 247–262.
- [15] S. DEMPE, *Foundations of bilevel programming*, Springer Science & Business Media, 2002.
- [16] S. DEMPE AND J. DUTTA, *Is bilevel programming a special case of a mathematical program with complementarity constraints?*, Mathematical programming, 131 (2012), pp. 37–48.
- [17] M. FUKUSHIMA AND J.-S. PANG, *Convergence of a smoothing continuation method for mathematical programs with complementarity constraints*, in Ill-posed Variational Problems and Regularization Techniques, Springer, 1999, pp. 99–110.
- [18] J.-B. HIRIART-URRUTY, *Refinements of necessary optimality conditions in nondifferentiable programming i*, Applied mathematics and optimization, 5 (1979), pp. 63–82.
- [19] C. KANZOW AND A. SCHWARTZ, *The price of inexactness: convergence properties of relaxation methods for mathematical programs with complementarity constraints revisited*, Mathematics of Operations Research, 40 (2014), pp. 253–275.

- [20] A. KESHAVARZ, Y. WANG, AND S. BOYD, *Imputing a convex objective function*, in Intelligent Control (ISIC), 2011 IEEE International Symposium on, IEEE, 2011, pp. 613–619.
- [21] W. KRICHENE, J. D. REILLY, S. AMIN, AND A. M. BAYEN, *Stackelberg routing on parallel networks with horizontal queues*, IEEE Trans. Automat. Contr., 59 (2014), pp. 714–727.
- [22] A. B. LEVY, R. POLIQUIN, AND L. THIBAUT, *Partial extensions of Attouch's theorem with applications to proto-derivatives of subgradient mappings*, Trans. Amer. Math. Soc., 347 (1995), pp. 1269–1294.
- [23] M. LIGNOLA AND J. MORGAN, *Topological existence and stability for stackelberg problems*, Journal of Optimization Theory and Applications, 84 (1995), pp. 145–169.
- [24] G.-H. LIN AND M. FUKUSHIMA, *A modified relaxation scheme for mathematical programs with complementarity constraints*, Annals of Operations Research, 133 (2005), pp. 63–84.
- [25] G.-H. LIN, M. XU, AND J. YE, *On solving simple bilevel programs with a nonconvex lower level program*, Mathematical Programming, 144 (2014), pp. 277–305.
- [26] P. LORIDAN AND J. MORGAN, *New results on approximate solution in two-level optimization*, Optimization, 20 (1989), pp. 819–836.
- [27] A. V. D. MIGUEL, M. P. FRIEDLANDER, F. J. NOGALES MARTÍN, AND S. SCHOLTES, *An interior-point method for MPECs based on strictly feasible relaxations.*, tech. report, Department of Decision Sciences, London Business School, 2004.
- [28] A. NEMIROVSKI, *Interior point polynomial time methods in convex programming*, tech. report, Georgia Institute of Technology, 2004.
- [29] Y. NESTEROV AND A. NEMIROVSKII, *Interior-Point Polynomial Algorithms in Convex Programming*, Society for Industrial and Applied Mathematics, 1994.
- [30] J. V. OUTRATA, *On the numerical solution of a class of Stackelberg problems*, Zeitschrift für Operations Research, 34 (1990), pp. 255–277.
- [31] E. POLAK, *Optimization: algorithms and consistent approximations*, vol. 124, Springer Science & Business Media, 1997.
- [32] R. POLIQUIN AND R. ROCKAFELLAR, *Prox-regular functions in variational analysis*, Transactions of the American Mathematical Society, 348 (1996), pp. 1805–1838.
- [33] R. A. POLIQUIN, *An extension of Attouch's theorem and its application to second-order epidifferentiation of convexly composite functions*, Trans. Amer. Math. Soc., 332 (1992), pp. 861–874.
- [34] R. ROCKAFELLAR, *Favorable classes of Lipschitz continuous functions in subgradient optimization*, in Progress in Nondifferentiable Optimization, E. Nurminski, ed., IIASA, Laxenburg, Austria, 1982, pp. 125–143.
- [35] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, 1970.
- [36] R. T. ROCKAFELLAR AND R. J.-B. WETS, *Variational analysis*, Springer, 3rd ed., 2009.
- [37] T. ROUGHGARDEN, *The price of anarchy is independent of the network topology*, Journal of Computer and System Sciences, 67 (2003), pp. 341–364.
- [38] J. O. ROYSET AND R. J. WETS, *Optimality functions and lopsided convergence*, Journal of Optimization Theory and Applications, (2016), pp. 1–19.
- [39] H. SCHEEL AND S. SCHOLTES, *Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity*, Math. Oper. Res., 25 (2000), pp. 1–22.
- [40] S. SCHOLTES, *Convergence properties of a regularization scheme for mathematical programs with complementarity constraints*, SIAM Journal on Optimization, 11 (2001), pp. 918–936.
- [41] Y. SHARMA AND D. P. WILLIAMSON, *Stackelberg thresholds in network routing games or the value of altruism*, in Proceedings of the 8th ACM conference on Electronic commerce, ACM, 2007, pp. 93–102.
- [42] S. STEFFENSEN AND M. ULBRICH, *A new relaxation scheme for mathematical programs with equilibrium constraints*, SIAM Journal on Optimization, 20 (2010), pp. 2504–2539.
- [43] C. SWAMY, *The effectiveness of stackelberg strategies and tolls for network congestion games*, ACM Transactions on Algorithms (TALG), 8 (2012), p. 36.
- [44] G. WACHSMUTH, *On LICQ and the uniqueness of Lagrange multipliers*, Operations Research Letters, 41 (2013), pp. 78–80.
- [45] J. YE AND D. ZHU, *Optimality conditions for bilevel programming problems*, Optimization, 33 (1995), pp. 9–27.
- [46] X. Y. ZHENG AND Z. WEI, *Convergence of the associated sequence of normal cones of a Mosco convergent sequence of sets*, SIAM Journal on Optimization, 22 (2012), pp. 758–771.
- [47] T. ZOLEZZI, *Continuity of generalized gradients and multipliers under perturbations*, Mathematics of Operations Research, 10 (1985), pp. 664–673.